Proceedings of the
22nd European Conference on
Cyber Warfare and Security

A Conference hosted by

Hellenic Air Force Academy and
the University of Piraeus, Greece

22-23 June 2023

Edited by
Professor Antonios Andreatos
and Professor Christos Douligeris

A conference managed by ACI, UK

aci

Proceedings of the

22nd European Conference on Cyber Warfare and Security

ECCWS 2023

A Conference Hosted By

Hellenic Air Force Academy (HAFA)
and
the University of Piraeus, Greece

22-23 June 2023

Copyright the authors, 2023. 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 Thomson ISI Web of Science and Elsevier Scopus.

Author affiliation details in these proceedings have been reproduced as supplied by the authors themselves.

The Electronic version of the Conference Proceedings is available to download from https://papers.academic-conferences.org/

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

E-Book ISBN: 978-1-914587-70-2
E-Book ISSN: 2048-8610
Book version ISBN: 978-1-914587-69-6
Book Version ISSN: 2048-8602

Published by Academic Conferences International Limited
Reading, UK
+44 (0) 118 324 6938
www.academic-conferences.org
info@academic-conferences.org

Contents

Preface ix

Committee x

Biographies xi

Academic Papers

A Survey on National Cyber Emergency Plans
Konstantinos Adamos, Ioannis Filippopoulos, George Stergiopoulos, Dimitris Gritzalis 1-11

Digital Forensic in A Virtual World; A Case of Metaverse and VR
tayba al ali, Sara Al Fulaiti, Manal Abuzour, Sheikha Almaqahami, Richard Ikuesan 12-21

Attention-Based Deep Learning Modelling for Intrusion Detection
Ban AlOmar, Dr. Zouheir Trabelsi, Dr. Firas Saidi 22-32

Towards the Development of Indicators of Fake Websites for Digital Investigation
Aysha Alkuwaiti, Mera Alremeithi, Haya Alobeidli, Richard Ikuesan 33-43

An Educational Scenario for Teaching Cyber Security Using low-cost Equipment and Open Source Software
Antonios Andreatos 44-53

A Methodical Framework for Conducting Reconnaissance and Enumeration in the Ethical Hacking Lifecycle
Fouz Barman, Nora Alkaabi, Hamda Almenhali, Mahra Alshedi, Richard Ikuesan 54-64

A Commentary and Exploration of Maritime Applications of Biosecurity and Cybersecurity Intersections
Michaela Barnett, Issah Samori, Brandon Griffin, Xavier-Lewis Palmer, Lucas Potter 65-72

Functional Architectural Design of a Digital Forensic Readiness Cybercrime Language as a Service
Stacey O Baror, Richard Adeyemi, I, Hein. S Venter 73-82

Teaching Social Science Aspects of Information Security
Igor Bernik 83-88

An Analysis of the MTI Crypto Investment Scam: User Case
Johnny Botha, Thor Pederson, Louise Leenen 89-99

How to safely communicate with a phishing attacker by email?
Ladislav Buřita, Aneta Coufalikova, Kamil Halouzka 100-107

Agile Methods For Improved Cyber Operations Planning
Jami Carroll 108-115

A New Interpretation of Integrated Deterrence: Physical and Virtual Strategies
Jim Chen 116-123

Influence Diagrams in Cyber Security: Conceptualization and Potential Applications
Sabarathinam Chockalingam, Clara Maathuis 124-131

Permission-Based Classification of Android Malware Applications Using Random Forest
Nikolaos Chrysikos, Panagiotis Karampelas, Konstantinos Xylogiannopoulos 132-142

Cultural Influences on Information Security
Henry Collier, Charlotte Morton, Dalal Alharthi, Jan Kleiner 143-150

Designing a high-fidelity testbed for 5G-based Industrial IoT
Diogo Cruz, Tiago Cruz, Vasco Pereira, Paulo Simões 151-160

Cognitive Security: Facing Cognitive Operations in Hybrid Warfare1
Didier Danet 161-168

JTF-ARES as a Model of a Persistent, Joint Cyber Task Force
Charlotte Donnelly, Marcel Stolz 169-176

Cyber power in the African context: an exploratory analysis and proposition
Petrus Duvenage, Wilhelm Bernhardt, Sebastian von Solms 177-186

DPIA for Cloud-based Health Organizations in the context of GDPR
Dimitra Georgiou, Costas Lambrinoudakis 187-198

Radiograph Manufacturer and Model Identification Using Deep-RSI
Farid Ghareh Mohammadi, Ronnie Sebro 199-206

Cyberspace Geography and Cyber Terrain: Challenges Producing a Universal map of Cyberspace
Alexander Grandin 207-213

Detect, Deny, Degrade, Disrupt, Destroy, Deceive: which is the greatest in OCO?
Tim Grant 214-222

Known Unknowns: The Inevitability of Cyber Attacks
Virginia Greiman 223-231

Complicity in Unlawful Offensive Cyber Operations Under International Law on State Responsibility
Samuli Haataja 232-238

Semiotics of Strategic Narratives of "Antichrist" in Russia's War in Ukraine
Michael Hotchkiss 239-247

Processing Model and Classification of Cybercognitive Attacks: Based on Cognitive Psychology
Ki Beom Kim, Eugene Lim, Hun Yeong Kwon 248-256

Designing an Email Attack by Analysing the Victim's Profile. An Alternative Anti-Phishing Training Method
Dimitrios Lappas, Panagiotis Karampelas 257-266

Designing Security for the Sixth Generation: About Necessity, Concepts and Opportunities
Christoph Lipps, Annika Tjabben, Matthias Rüb, Jan Herbst, Sogo Pierre Sanon,
Rekha Reddy, Yorman Munoz, Hans D. Schotten 267-275

Tackling Uncertainty Through Probabilistic Modelling of Proportionality in Military Operations
Clara Maathuis, Sabarathinam Chockalingam 276-284

Design Lessons from Building Deep Learning Disinformation Generation and Detection Solutions
Clara Maathuis, Iddo Kerkhof, Rik Godschalk, Harrie Passier 285-293

Teaching pentesting to social sciences students using experiential learning techniques to improve attitudes towards possible cybersecurity careers
Aleksandras Melnikovas, Ricardo G. Lugo, Kaie Maennel, Agnė Brilingaitė,
Stefan Sütterlin, Aušrius Juozapavičius 294-302

Spam Email Detection Using Machine Learning Techniques
Mr, Antonios Andreatos, Prof 303-310

Cyber Warfare and Cyber Terrorism Threats Targeting Critical Infrastructure: A HCPS-based Threat Modelling Intelligence Framework
Rennie Naidoo, Carla Jacobs 311-318

Determination of the end device risk likelihood using the Bayesian network tools
Tabisa Ncubukezi 319-331

Enabling fine-grained access control in information sharing with structured data formats
Tatu Niskanen, Jarno Salonen 332-340

Towards an active cyber defence framework for SMMEs in developing countries
Nombeko Ntingi, Prof Sebastian von Solms, Dr Jaco du Toit 341-348

Participants Prefer Technical Hands-on Cyber Exercises Instead of Organisational and Societal Ones
Jani Päijänen, Jarno Salonen, Anni Karinsalo, Tuomo Sipola, Tero Kokkonen 349-357

A Reflection on Typology and Verification Flaws in consideration of Biocybersecurity/Cyberbiosecurity: Just Another Gap in the Wall
Luke Potter, Kim Mossberg, Xavier Palmer 358-365

Assessment of Cyber Security risks: A Smart Terminal Process
Jouni Pöyhönen 366-373

Governance and management information system for cybersecurity centres and competence hubs
Jyri Rajamäki, Janne Lahdenperä 374-383

Students' Application of the MITRE ATT&CK® Framework via a real-time Cybersecurity Exercise
Aunshul Rege, Jamie Williams, Rachel Bleiman, Katorah Williams 384-394

Role of Techno-Economic Coalitions in Future Cyberspace Governance: 'Backcasting' as a Method for Strategic Foresight
Mari Ristolainen 395-402

Digital Streets of Rage: Identifying Rhizomatic Extremist Messages During a Hybrid Media Event using Natural Language Processing
Teija Sederholm, Petri Jääskeläinen, Milla Lonka, Aki-Mauri Huhtinen 403-409

An Analysis of Critical Cybersecurity Controls for Industrial Control Systems
Nkata Sekonya, Siphesihle Sithungu 410-419

NCSS: A global census of national positions on conflict, neutrality and cooperation
Radu Antonio Serrano Iova, Tomoe Watashiba 420-428

Developing Cybersecurity in an Industrial Environment by Using a Testbed Environment
Jussi Simola, Reijo Savola, Tapio Frantti, Arttu Takala, Riku Lehkonen 429-438

Smart Terminal System of Systems' Cyber Threat Impact Evaluation
Jussi Simola, Jouni Pöyhönen, Martti Lehto 439-449

A Cyber Counterintelligence Competence Framework: Developing the Job Roles
Thenjiwe Sithole, Jaco Du Toit, Sebastian von Solms 450-457

Static Vulnerability Analysis Using Intermediate Representations: A Literature Review
Adam Spanier, William Mahoney 458-465

Legal and ethical issues of pre-incident forensic analysis.
Dr Iain Sutherland, Dr Matthew Bovee, Dr Konstantinos Xynos, Dr Huw O. L. Read 466-473

Developing Robust Cyber Warfare Capabilities for the African Battlespace
Jabu Mtsweni, Mphahlela 474-483

Cybersecurity Through Thesis in Laurea University of Applied Sciences
Ilona Frisk, Harri Ruoslahti, Ilkka Tikanmäki 484-492

On the software architectures for fog-based secure IOT deployments
christos tselikis 493-499

Towards Norms for State Responsibilities regarding Online Disinformation and Influence Operations
Brett van Niekerk, Trishana Ramluckan 500-509

The Identification of Cybersecurity Work Roles for the Water Sector in South Africa
Sune von Solms 510-516

Cyber Lessons that the World Can Learn from Lithuania
Matthew Warren, Darius Štitilis, Marius Laurinaitis 517-524

Legal Response to Social Media Disinformation on National Level
Murdoch Watney 525-532

On Benchmarking and Validation in Wargames
Adam Wilden, Mehwish Nasim, Peter Williams, Tim Legrand,
Benjamin Peter Turnbull, Patricia A. H. Williams 533-543

The UN Global Digital Compact (GDC), Achieving a trusted, free, open, and Secure Internet: Trust-building
Allison Wylde 544-551

PhD Papers

Cybersecurity in Mozambique: Status and Challenges
Martina De Barros 553-558

Spreading Lies Through the Cyber Domain
Thomas Dempsey 559-566

The Concept of Comprehensive Security as a Tool for Cyber Deterrence
Maria keinonen 567-574

Building Situational Awareness of GDPR
Pauliina Hirvonen, Martti J. Kari 575-583

Organisational GDPR Investments and Impacts
Pauliina Hirvonen 584-591

A Whole-of-Society Approach to Organise for Offensive Cyberspace Operations: The Case of the Smart State Sweden
Gazmend Huskaj, Stefan Axelsson 592-602

A State-of-the-art of Scientific Research on Disinformation
Gazmend Huskaj, Stefan Axelsson 602-609

Fake news as a distortion of media reality: tell-truth strategy in the post-truth era
Anastasiia Iufereva 610-615

Hybrid threats-possible consequences in societal contexts
Georgiana Daniela Lupulescu 616-622

Target Audiences' Characteristics and Prospective in Countering Information Warfare
Daniel Ionel Andrei Nistor 623-630

Technology-Oriented Innovations and Cyber Security Challenges in the Healthcare Delivery System: A perspective from Developing Economy
Victor Kwarteng Owusu, Professor Gregar 631-638

Deep-learning-based Intrusion Detection for Software-defined Networking Space Systems
Uakomba Uhongora, Ronald, Law, Jill *639-647*

Masters Papers

How Does the Tallinn Manual 2.0 Shed Light on the Threat of Cyber Attacks against Taiwan?
Chih-Hsiang Chang 649-656

Digital Forensic Readiness Model for Internet Voting
Edmore Muyambo, Stacey O. Baror 657-667

Non-Academic Papers

We see what we want to see: Pitfalls of Perception and Decision-making in Security Management
Helvi Salminen 669-677

Zero Trust: The Magic Bullet or Devil's Advocate?
Helvi Salminen 678-686

Work in Progress Papers

Reconnaissance Techniques and Industrial Control System Tactics Knowledge Graph
Thomas Heverin 688-695

AI-based quantum-safe cybersecurity automation and orchestration for edge intelligence in future networks
Aarne Hummelholm 696-702

Security Issues of GPUs and FPGAs for AI-powered near & far Edge Services
Stylianos Koumoutzelis, Ioannis Giannoulakis, Titos Georgoulakis,
George Avdikos, Emmanouil Kafetzakis 703-706

Hybrid Threat and Information Influence in Connection with Security of Supply
Jyri Rajamäki, Tehi Palletvuori 703-710

Demand Analysis of the Cybersecurity Knowledge Areas and Skills for the Nurses: Preliminary Findings
Jyri Rajamäki, Paresh Rathod, Kitty Kioskli 711-716

Hidden Permissions on Android: A Permission-Based Android Mobile Privacy Risk Model
Saliha Yilmaz, Mastaneh Davis 717-724

Late Submission

From Provoking Emotions to fake Images: The Recurring Signs of fake news and Phishing Scams Spreading on Social Media in Hungary, Romania and Slovakia
Kenyeres Attila Zoltán and Lauren Weigand 726-732

Preface

These proceedings represent the work of contributors to the 22nd European Conference on Cyber Warfare and Security (ECCWS 2023), which is held in Athens, Greece on 22 and 23 June 2023. ECCWS 2023 has been supported by the Hellenic Air Force Academy and the University of Piraeus. The Conference Chair is Professor Antonios Andreatos from the Hellenic Air Force Academy, Greece and the Programme Chair is Professor Christos Douligeris from the University of Piraeus, Greece.

ECCWS is a well-established event on the academic research calendar and now in its 22nd year its key aim remains the opportunity for participants to network and share ideas. The scope of the papers ensure an interesting conference. 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. Dr. Kai Rannenberg, from Goethe University Frankfurt, Germany on the topic of *Multilateral Security: Defending one's assets without destroying one's values.* The second day of the conference will open with an address by Prof. Nineta Polemi, University of Piraeus, Greece speaking on *CyberSecPro: An agile and collaborative approach to cybersecurity practical trainings.*

With an initial submission of 132 abstracts, after the double blind peer review process 64 academic research papers, 12 PhD research papers, 2 Masters research papers 2 Non-Acdemic research papers and 5 work-in-progress papers are published in the Conference Proceedings. These papers represent research from Australia, Austria, Czech Republic, Estonia, Finland, France, Germany, Greece, Italy, Lithuania, Malaysia, Mozambique, Netherlands, Nigeria, Norway, Portugal, Republic of Korea, Romania, Slovenia, South Africa, Sweden , Switzerland, Taiwan, UK, United Arab Emirates and the USA.

We hope that you will enjoy the conference and that you will find the opportunity to renew past collaborations and friendships and to engage in fruitful discussions leading to new advancements in research and development.

Professor Antonios Andreatos
Hellenic Air Force Academy
Professor Christos Douligeris
University of Piraeus
Greece
June 2023

ECCWS Conference Committee

Dr Mohd Faizal Abdollah, University Technical Malaysia Melaka, Malaysia; Dr William ("Joe") Adams, University of Michigan/Merit Network, USA; Prof Hamid Alasadi, Basra University, Iraq; Dr Kari Alenius, University of Oulu, Finland; Prof Antonios Andreatos, Hellenic Air Force Academy, Greece; Dr Olga Angelopoulou, University of Warwick, UK; John Anohar, Higher Education dept, India; Dr Leigh Armistead, Edith Cowan University, Australia; Dr Sajad Bagheri Baba Ahmadi, , Iran; Dr Hayretdin Bahsi, Tallinn University of Technology, Estonia; Dr Dragos Barbieru, CAROL I National Defence University, Romania; Prof Jorge Barbosa, Coimbra Polytechnic - ISEC, Portugal; Stacey Baror, University of Pretoria, South Africa; Prof Matt Bishop, University of California at Davis, USA; Dr Jim Chen, DoD National Defense University, USA; Dr Sabarathinam Chockalingam, Institute for Energy Technology, Norway; Dr Maura Conway, Dublin City University, Ireland; Dr Paul Crocker, Universidade de Beira Interior, Portugal; Prof Tiago Cruz, University of Coimbra, Portugal; Dr Christian Czosseck, CERT Bundeswehr (German Armed Forces CERT), Germany; Josef Demergis, University of Macedonia, Greece; Prof Patricio Domingues, Polytechnic Institute of Leiria, Portugal; Paul Dowland, Edith Cowan University, Australia; Marios Efthymiopoulos, Political Science Department University of Cyprus, Cyprus; Dr Colin Egan, University of Hertfordshire, Hatfield, UK; Dr Ruben Elamiryan, Public Administration Academy of the Republic of Armenia, Armenia; Dr Thaddeus Eze, University of Chester, UK; Prof Eric Filiol, ENSIBS, Vannes, France & CNAM, Paris, France; Prof Steve Furnell, University of Nottingham, UK; Dr Dimitra Georgiou, University of Piraeus, Greece; Dr Michael Grimaila, Air Force Institute of Technology, USA; Bill Hutchinson, Edith Cowan University, Australia; Hamid Jahankhani, University of East London, UK; Nor Badrul Anuar Jumaat, University of Malaya, Malaysia; Dr Connie Justice, Purdue School of Engineering and Technology, IUPUI, USA; Maria Karyda, University of the Aegean, Greece; Anthony Keane, Technological University Dublin, Ireland; Jyri Kivimaa, Cooperative Cyber Defence and Centre of Excellence, Tallinn, Estonia; Prof Mortaza Kokabi, Shaheed Chamran Univ., Iran; Prof Ahmet Koltuksuz, Yasar University, Dept. of Comp. Eng, Turkey; Peter Kunz, Hexagon Geosystems, Germany; Martti Lehto, National Defence University, Finland; Christoph Lipps, German Research Center for Artificial Intelligence, Germany; Dr Efstratios Livanis, University of Macedonia, Greece; Dr Leandros Maglaras, De Montfort University, UK; Assc Manmeet Mahinderjit Singh , University Sains Malaysia, Malaysia; James Malcolm, University of Hertfordshire, UK; Dr Arianit Maraj, AAB College-Faculty of Computer Sciences, Kosovo; Ioannis Mavridis , University of Macedonia, Greece; Rob McCusker, Charles Sturt University, UK; Dr Yonathan Mizrachi, University of Haifa, Israel; Antonio Muñoz, University of Málaga, Spain; Daniel Ng, C-PISA/HTCIA, China; Dr Cyril Onwubiko, Cyber Security Intelligence at Research Series Limited, UK; Rain Ottis, Tallinn University of Technology, Estonia; Prof Teresa Pereira, Instituto Politécnico de Viana do Castelo, Portugal; Michael Pilgermann, Technische Hochschule Brandenburg - University of Applied Sciences, Germany; Prof Pedro Pinto, Instituto Politécnico de Viana do Castelo, Portugal; Dr Ajeet Singh Poonia, Full time academic, Bikaner Technical University, Rajasthan; Prof Maria-Magdalena Popescu, Carol I National Defence University , Romania; Dr Bernardi Pranggono, Sheffield Hallam University, UK; Prof Carlos Rabadão, Politechnic of Leiria, Portugal; Dr Trishana Ramluckan, University of KwaZulu-Nata, South Africa; Dr Mark Reith, Full time academic, Air Force Institute of Technology; Dr Neil Rowe, US Naval Postgraduate School, Monterey, USA; Prof Vitor Sa, Catholic University of Portugal, Portugal; Dr Char Sample, Carnegie Mellon University/CERT, USA; Prof Henrique Santos, University of Minho, Portugal; Prof Leonel Santos, Polytechnic of Leiria, Portugal; Dr Keith Scott, De Montfort University, UK; Prof Dr Richard Sethmann, University of Applied Sciences Bremen, Germany; Dr Yilun Shang, Northumbria University, UK; Prof Jill Slay, University of South Australia, Australia; Dr Lee Speakman, University of Chester, UK; Dr Joseph Spring, University of Hertfordshire, UK; Prof Iain Sutherland, Noroff University College , Norway; Dr Hamed Taherdoost, Hamta Group, Hamta Business Corp, Vancouver, Canada; Unal Tatar, University at Albany - SUNY, USA; Dr Selma Tekir, Izmir Institute of Technology, Turkey; Prof Dr Peter Trommler, Georg Simon Ohm University Nuremberg, Germany; Mr Paul Underhill, University of Chester, UK; Dr Brett van Niekerk, Durban University of Technology, South Africa; Richard Vaughan, General Dynamics UK Ltd, UK; Dr Namosha Veerasamy, Council for Scientific and Industrial Research, South Africa; Prof Trish Williams, Flinders University, Australia; Prof Richard Wilson, Towson University, USA; Dr Allison Wylde, Glasgow Caledonian University , UK; Dr Justyna Żywiołek, Czestochowa University of Technology, Poland.

Biographies

Conference and Programme Chairs

Antonios Andreatos is a professor at the Computer Engineering Division of the Hellenic Air Force Academy. He serves as a scientific committee member for a number of international conferences, as well as a reviewer in several journals. He has published papers in various international journals and conference proceedings, book chapters and three books. His research interests include (among other topics) cyber security, cryptography, steganography, image processing, educational technology, IoT, etc.

Christos Douligeris, currently a professor at the Department of Informatics, University of Piraeus, Greece, held positions with the Department of Electrical and Computer Engineering at the University of Miami. He was an associate member of the Hellenic Authority for Information and Communication Assurance and Privacy and the President and CEO Hellenic Electronic Governance for Social Security SA. Dr. Douligeris has published extensively in the networking scientific literature and he has participated in many research and development projects. His main research interests lie in the areas of computer networking, communications, network security, cyber security, web science, data analytics, new technologies in education and emergency response operations.

Keynote Speakers

Nineta Polemi is a cybersecurity Professor in the University of Piraeus-UNIPI- (Cyber Security Lab, Dept. of Informatics) and CTO/ Co-Founder of trustilio. She established and acted as director of the UNIPI cybersecurity lab. She served (2017-2020) as Programme Manager and Policy Officer in the European Commission DG (CONNECT H1 Unit entitled 'Cybersecurity Technologies and Capabilities'). Her cybersecurity expertise is on: operational risk/conformity assessment, security management, incident management, supply chain security, threat intelligence. She has obtained her Ph.D. in Applied Mathematics (Coding Theory) from The City University of New York (Graduate Center). She held teaching and research positions in The City University of New York (Queens & Baruch Colleges), State University of New York (Farmingdale), Université Libre de Bruxelles (ULB)-Solvay Brussels School-. She has over 150 publications in security has organised numerous scientific and policy international cybersecurity scientific events and trainings e.g. cyberhot. She has received many research grants (NATO, IEEE) and awards (NSA, MSI Army Research Office IEEE, CYNY, Hellenic Ministry of Maritime, Hellenic National Defense General) and has participated as Project and Technical Manager in more than 60 cybersecurity international, EU and national R&D and commercial projects. She serves as external expert/reviewer/consultant in EC, ENISA, FORTH, Focal Point, CodeWeTrust. She is a member of ETSI in the working groups of Cyber and AI.

Kai Rannenberg, Chair of Mobile Business & Multilateral Security (www.m-chair.de) at Goethe University Frankfurt since 2002. Visiting Professor National Institute for Informatics (NII) (Tokyo, Japan) since 2012. Chair CEPIS (www.cepis.org) Legal & Security Issues Special Interest Network since 2003. Since 2007 Convenor ISO/IEC JTC 1/SC 27/WG 5 "Identity management & privacy technologies". 2004-2013 academic expert in the Management Board of ENISA; 2013-2020 member ENISA's Advisory Group (until 2019 Permanent Stakeholder Group). Since 2021 Honorary Treasurer International Federation for Information Processing, 1999-2002 Microsoft Research Cambridge (UK). Kai has been coordinating several leading EU research projects, e.g. the Network of Excellence "Future of Identity in the Information Society (FIDIS)" and the IP "Attribute based Credentials for Trust" (ABC4Trust).

Currently he is coordinating CyberSec4Europe, a pilot for the Cybersecurity Competence Network the EU is aiming for, and CyberSecPro aiming for improved cybersecurity education and training.

Mini Track Chairs

Olga Angelopoulou is an Assistant Professor in Cyber Security at the University of Warwick. Olga completed her doctorate in digital forensics in 2010 at the University of South Wales (Glamorgan) and is an active researcher in the field. She previously held academic posts at the University of Derby and the University of Hertfordshire. Olga is also the Editor-in-Chief for Information Security Journal: A global Perspective. Her research interests include digital investigations, cybercrimes, incident response and online fraud. She has several publications and presentations in the field and is involved in numerous scholarly activities.

Sabarathinam Chockalingam is a Research Scientist I at the Institute for Energy Technology in Halden, Norway. Saba has a PhD in Cyber Security and Artificial Intelligence from the Delft University of Technology and MSc in Cyber Security and Management from the University of Warwick. His research interests include cyber security, risk management and safety.

Tiago Cruz received his Ph.D. degree in informatics engineering from the University of Coimbra (Coimbra, Portugal), in 2012. He is an Assistant Professor in the Department of Informatics Engineering, University of Coimbra. His research interests include areas such as management systems for communications infrastructures and services, critical infrastructure security, broadband access network device and service management, Internet of Things.

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.

Clara Maathuis is an Assistant Professor in AI and Cyber Security at Open University in the Netherlands. With a PhD in AI and Military Cyber Operations from the Delft University of Technology, she is involved in teaching different AI and cyber security courses and conducts research in AI, cyber/information operations, military technologies, and social manipulation.

Shahzad Ahmed Memon is currently working as senior researcher at department of computer Engineering, Faculty of Engineering and architecture in Beykoz University, Istanbul, Turkey. He also worked as Professor, at department of Electronics Engineering, Faculty of Engineering and Technology at University of Sindh, Pakistan. He completed his Doctorate PhD in Biometrics security from Brunel University, London, UK. He published his research in several national and international research journals. Dr. Memon attended and presented his research in national and international conferences organized in USA, UK, and Europe. Dr. Memon supervised MS and PhD students in the field of Biometrics, Cyber Security and Privacy, Smarts systems security, Cyber Physical systems security, Applications of Artificial

Intelligence. He also granted funding for research from Higher Education commission and ICT Research and Development, Ministry of Information Technology, Pakistan

Brett van Niekerk is a senior lecturer in information technology at the Durban University of Technology. He serves as chair for the International Federation of Information Processing Working Group on ICT in Peace and War, and the co-Editor-in-Chief of the International Journal of Cyber Warfare and Terrorism. He has numerous years of information/cyber-security experience in both academia and industry, and has contributed to the ISO/IEC information security standards. In 2012 he graduated with his PhD focusing on information operations and critical infrastructure protection. He is also holds a MSC in electronic engineering and is CISM certified.

Trishana Ramluckan is a Honorary Researcher in the School of Law at the University of KwaZulu-Natal and Research Manager at Educor Holdings. She is a member of the IFIP working group on ICT Uses in Peace and War, the Institute of Information Technology Professionals South Africa and is an Academic Advocate for ISACA. In 2017 she graduated with a Doctor of Administration specialising in IT and Public Governance and in 2020 she was listed as in the Top 50 Women in Cybersecurity in Africa. Her current research areas include Cyber Law and Information Technology Governance.

Dr Keith Scott Is Programme Leader for English Language at De Montfort University in Leicester. His research operates at the intersection of communication, culture and cyber, with particular interests in influence, information warfare, and simulations and serious gaming as a training. teaching, and research tool. He is a member of the Prometheus Research group, supported by the UK Development, Concepts and Doctrine Centre, which investigates the application of cybernetics and systems thinking within the context of warfare, as well as more general considerations of AI and pervasive technology.

Paulo Simões received his Ph.D. degree in informatics engineering from the University of Coimbra (Coimbra, Portugal), in 2002. He is an Associate Professor in the Department of Informatics Engineering, University of Coimbra. His research interests include network and infrastructure management, security, critical infrastructure protection.

Workshop Facilitators

Dr Edwin "Leigh" Armistead is the Principal of ArmisteadTec LLC, a Veteran-Owned Small Business (VOSB) that is certified by the Small Business Administration (SBA) as a Historically Underutilized Business Zone (HBZ), https://www.armisteadtec.com/. This company primarily focuses on the academic functions of Information Warfare (IW), to include curriculum development, teaching as well as other related tasks to include lecturing and writing. A retired United States Naval Officer, Dr Armistead has significant Information Operations academic credentials having written his PhD on the conduct of Cyber Warfare by the federal government and has published three books, in an unclassified format in 2004, 2007 and 2010, all focusing on full Information Warfare. He is also the Chief Editor of the Journal of Information Warfare (JIW) https://www.jinfowar.com/; the Program Director of the International Conference of Cyber Warfare and Security and the Vice-Chair Working Group 9.10, ICT Uses in Peace and War. Shown below are the books on full spectrum cyber warfare and the JIW:

Biographies of Contributing Authors

Konstantinos Adamos (B.Sc., M.Sc.) Cybersecurity Researcher with the Information Security and Critical Infrastructure Protection Research Group (Athens University and Economics & Business, Greece). Also a Ph.D. candidate with the Dept. of Information and Communication Systems Engineering (University of the Aegean, Greece). His current research interests focus on cybersecurity, critical infrastructure protection, data protection, and risk assessment.

Tayba Al Ali is a graduate of Information technology with specialization in cyber security and networking at Zayed University. She's certified in Introduction to Networks and Switching, Routing, and Wireless Essentials, Tayba's constant appearance on the dean's list during her studies at the University is a testament to her academic prowess. Her passion for technology guarantees a successful career in IT.

Nora AlKaabi is an Information Technology graduate from Zayed University, specializing in Security and Network Technology. Her fascination with technology and cybersecurity fuels her drive to create innovative solutions for businesses to overcome security challenges. Nora is passionate about contributing to a secure and digitally-driven society, utilizing her skills and knowledge to make a positive impact on the world.

Ban AlOmar is currently pursuing her Ph.D. degree with the College of Information Technology, United Arab Emirates University. Her main research areas are Artificial Intelligence, machine learning, network security and intrusion detection and prevention.

Antonios S. Andreatos is Professor at the Hellenic Air Force Academy, Greece. He has published his research in international journals, conference proceedings and book chapters. He is involved in scientific committees of several conferences and reviewer in several journals. His current research interests focus on Cryptography, Image processing, CyberSecurity, IoT, Engineering Education, e-Assessment, CyberSecurity Education.

Stacey Omeleze Baror obtained her B.sc Hon and Master in Computer Sci., From the Computer Science Department, University of Pretoria. She is currently studying towards a PhD in Computer Science, at the University of Pretoria in research areas focusing on digital forensic, Software engineering, intelligent systems, Digital Forensics Application, Digital Forensic Readiness, and Information Security.

Igor Bernik is Professor of Information Sciences and vice dean for Quality and Development at the Faculty of Criminal Justice and Security, University of Maribor, Slovenia. His research fields are information systems, cybersecurity, and the growing requirement for cybersecurity awareness. He is an (co)author of several scientific books and articles published in recognized international journals.

Micki Boland is a global cybersecurity warrior and evangelist with Check Point Software Technologies Office of the CTO. A practitioner with 20 years in IT, cybersecurity, emerging technology innovation, Micki holds ISC2 CISSP, Master of Science in Technology Commercialization University of Texas at Austin and MBA Global Security Concentration from East Carolina University.

Johnny Botha enrolled for a PhD in Computer Science at the University of Western Cape. Topic: *"A process for Blockchain Crime Investigations Based on Open-source Intelligence".* Project Manager & Researcher (CSIR). Masters degree in Information Technology, at University of South Africa (UNISA). NDip and BTech degree in Computer Systems Engineering at the Tshwane University of Technology (TUT).

Ladislav Burita, University of Defence, Brno, Czech Republic. PhD. from 1985 and professor from 2003. He worked in NATO/MIP team; was head of the faculty research program and defense project MENTAL. His area of interest is information and knowledge systems. He published a hundred papers at conferences and in journals.

Jami M. Carroll retired from the Navy after 20 years. He later founded Prisidian Security Solutions in 2008. His research areas include offensive cyber operations, threat intelligence, denial/deception, and geopolitical instability. His degrees include: BBA, National University; MBA, Southern New Hampshire University; MS/DSc in in Cybersecurity, Capitol Technology University.

Chih-Hsiang ('Sharon') Chang is a Master student of Interdisciplinary Legal Studies at National Taiwan University. She is the co-author of The Best Memorial-Defendant of Red Cross International Humanitarian Law Moot 2021 Hong Kong. The main focus of her research is international law, law of armed conflict and cyber warfare.

Jim Q. Chen, Ph.D. is Professor and Associate Dean at the U.S. National Defense University College of Information and Cyberspace. His expertise is in cyber strategy, cyber deterrence, cyber warfare, cybersecurity technology, artificial intelligence, and machine learning. He has authored and published more than 90 peer-reviewed papers, articles, and book chapters on these topics. He has also been teaching graduate courses on these topics. He is a recognized expert in cyber studies and artificial intelligence.

Sabarathinam Chockalingam is a Research Scientist I at the Institute for Energy Technology in Halden, Norway. Saba has a PhD in Cyber Security from the Delft University of Technology and MSc in Cyber Security and Management from the University of Warwick. His research interests include cyber security, risk management, serious games.

Nikolaos Chrysikos is an electrical engineer in the Hellenic Air Force. He graduated from Hellenic Air Force Academy in 2022 . His diploma thesis was on malicious android applications analysis and classification. He is interested in data analysis and Cyber Security.

Henry Collier is an associate professor and the Director for the Technology Programs at Norwich University's college of Graduate and Continuing Studies. He received his Ph.D. in engineering security from the University of Colorado Colorado Springs. His main research areas are human factors in cybersecurity and networking.

Tiago Cruz is an Associate Professor with the Department of Informatics Engineering of the University of Coimbra. His research interests cover areas such as management systems for communications infrastructures and services, critical infrastructure security, broadband access network device and service management, Internet of Things, software defined networking, and network function virtualization.

Didier Danet is a senior lecturer of Defense management at Saint-Cyr Military Academy, France. He holds a PhD in Management Sciences from Rennes 1 University. He is the Director of the Post-Master Training Program “Cyberdefense and Intangible Fields”. His main research areas are Cyber Crisis Management, Military Organisations and Irregular Warfare.

Martina Jennifer Zucule de Barros is a PhD student in cybersecurity. Her research explores cybersecurity education. She has a master degree in Distributed Systems Engineering from the Technical University of Dresden. She is currently working as a university teacher in Computer Science at Universidade Pedagogica de Maputo (UPM).

Thomas Dempsey is a second-year Doctoral student in Strategic Intelligence. Thomas has a Masters degree in Cybersecurity and multiple industry certifications. In addition to academics, Thomas is a Cybersecurity professional who has worked in cyber for over 14 years, including time on active duty in the United States Marine Corps conducting defensive cyber operations.

Charlie Donnelly is studying Computer Science and Economics at the University of Chicago. She has worked with the Securities and Exchange Commission and Morgan Stanley’s Technology Investment Banking Group. She is interested in studying military applications of cybersecurity.

Petrus ('Beer') Duvenhage has extensive experience in the field of information security. He holds a PhD from the University of Pretoria and a PhD from the University of Johannesburg.

Bastian Fraune is a computer-scientist and works as a researcher in industrial cyber security projects. His research interests include secure architectures for critical infrastructures, cyber-physical systems and digitization of systems. He holds a master's degree in computer science and is working on secure power grid architectures in his PhD-project.

Dimitra Georgiou: Lecturer at the Computer Engineering Division of the Hellenic Air Force Academy and an Adjunct Lecturer (Department of Digital Systems, University of Piraeus). She received her PhD in Security Policies of Cloud Computing from the University of Piraeus in 2018. Research interests lie in the areas of information and communication Systems Security and Privacy.

Farid Gharehmohammadi is a Mayo Clinic postdoctoral researcher investigating deep learning in medical imaging. He earned his Ph.D. in computer science from University of Georgia. He has a decade of industrial and research experience. He published over 20 journal, conference papers and book chapters. He received the Outstanding Achievement Award for deep learning at CSCI’19.

Alexander Grandin has a Master’s degree in Science (2022) and Military science (2018) and works as a research officer at the Finnish National Defense University (FNDU). He has served in the Finnish Defense Forces from 2012.

Tim Grant is retired but an active researcher (Professor emeritus, Netherlands Defence Academy). Tim has a BSc in Aeronautical Engineering (Bristol University), a Masters-level Defence Fellowship (Brunel University), and a PhD in Artificial Intelligence (Maastricht University). Tim's research focuses on Offensive Cyber Operations, Command & Control, and Emergency Management. More details are at https://www.linkedin.com/in/tim-grant-r-bar/.

Virginia A. Greiman. Assistant Professor (Global Cyber Law, Governance, Megaprojects and Infrastructure Planning, Boston University). Was a diplomatic head of delegation (U.S. Department of State in Eastern and Central Europe, Asia and). Held several high-level appointments with the U.S. Department of Justice. Conducts research and publishes in cyber resilience, cyber law and regulation, nuclear cyber security, and global infrastructure megaprojects.

Samuli Haataja is a Senior Lecturer at Griffith Law School, Griffith University. He researches state-sponsored cyber-operations under public international law. He published his book 'Cyber Attacks and International Law on the Use of Force: The Turn to Information Ethics' with Routledge in 2019, and he has published in various international law and technology journals.

Thomas Heverin is the lead of Girls Learn Cyber, Inc., a training company for K-12 girls in cybersecurity. He holds a Ph.D. focused on cybersecurity, the CISSP, and years of cybersecurity experience. He led research and development projects for the U.S. Navy after serving as a Surface Warfare Officer."

Pauliina Hirvonen is a PhD student in the Information Technology Department of the University of Jyväskylä, in Finland. Her on-going dissertation of GDPR relates to Information Systems and especially on information and/or cyber security field. The research considers the overall impacts of privacy regulation and developing situational awareness from the organisational perspectives.

Michael Bennett Hotchkiss is an independent researcher with interests in the psychology of influence, criminology, and nation-state disinformation; especially in the context of Russian active measures. Michael earned a Master of Organization Development from Bowling Green State University; and a Bachelor of Arts in Psychology from the University of Connecticut, graduating Phi Beta Kappa.

Aki-Mauri Huhtinen (LTC (GS)) is a military professor at the Finnish National Defence University's Department of Leadership and Military Pedagogy. His areas of expertise are military leadership, command and control, information warfare, the philosophy of science in military organizational research.

Aarne Hummelholm, PhD in Information Technology (University of Jyväskylä, 2019). He has over 30 years' experience in the design, development of architectures` of authorities` telecommunications networks and information systems. Key themes in his work have been critical service availability, usability, cyber security and preparedness issues.

Gazmend Huskaj: Head of Cyber Security at GCSP. Was a doctoral student (Cyberspace Operations, Swedish Defence University.) Was Director of Intelligence on Cyber-related issues in the Swedish Armed Forces. A veteran with over five years of duty in conflict / post-conflict areas including two tours to the Balkans and one in Central Asia. Current work focuses on capacity-building, dialogue, and policy analysis.

Anastasiia Iufereva Ph.D. Student (IULM University, Milan, Italy). She has been awarded academic degrees: a bachelor's degree in journalism (2012), and a master's degree in management(2014). Her research interests oscillate between mass media and social

practices. In particular, she studies how journalists' professional cultures are shaped in the light of online platforms and media convergence.

Panagiotis Karampelas is an assistant professor at the Hellenic Air Force Academy, Greece. He received his PhD in electronic engineering from University of Kent in 2005. He participates in the advisory board of several scientific journals and conferences. His main research areas are Software Engineering, Data Mining, Cyber Security, Digital Forensics, Social Network Analysis

Major Maria Keinonen is a senior research officer and a PhD student for Military Sciences in the National Defence University in Finland. In her doctoral research, she focuses on the study of cyber deterrence from the perspective of a small state.

Stylianos Koumoutzelis is a graduate of the Electrical and Computer Engineering School of the National Technical University of Athens. With a focus on hardware design and development, he possesses expertise in the design and implementation of complex systems. He has also worked extensively in the fields of AI/ML, Telecommunications, and European research projects.

Eugene Lim is currently a master's student at Korea University's Graduate School of Cyber Security. Lim's bachelors degree is in International Studies and is currently interested in international cybersecurity and cyber policy. Her global research publications include 'Review of the Validity and Rationality of Artificial Intelligence Regulation' in the Florida Artificial Intelligence Research Society.

Georgiana-Daniela Lupulescu works at a public institution that offers specialization/improvement courses for military personnel. She is currently a PhD student at the National Defence University, conducting a research on the influence of hybrid threats in the realization of resilience policies in the national security field.

Clara Maathuis is Assistant Professor in AI and Cyber Security at Open University in the Netherlands. With a PhD in AI and Military Cyber Operations from Delft University of Technology, she is involved in teaching different AI and cyber security courses and conducts research in AI, cyber/information operations, military technologies, and social manipulation.

Bill Mahoney is the Charles and Margre Durham Distinguished Professor in Cybersecurity at the University of Nebraska at Omaha, where he teaches both undergraduate and graduate cybersecurity classes. His research areas include code obfuscation, reverse engineering and anti-reverse engineering techniques, and vulnerability analysis. An additional interest is in vulnerabilities with critical infrastructure equipment.

Aleksandras Melnikovas, Senior instructor at the Leadership and Methodology Development Centre, Military Academy of Lithuania. Obtained Masters degree in Education sciences in 2017, currently PhD student in Political sciences. Since 2021 works as a junior scientist within the project „Advancing Human Performance in Cybersecurity".

Jabu Mtsweni is the Head of Information and Cyber Security Centre and Chief Researcher at the Council for Scientific and Industrial Research (CSIR), NRF-Rated Researcher (C2), and a Certified Cybersecurity Manager, Research Fellow at the Stellenbosch University. He has

published over 80 peer-reviewed conference and journal articles over the years in different forums.

Edmore Muyambo is a Master's Research Student at the University of Pretoria, South Africa. He received his Honours Degree in Computing from the University of South Africa in 2022. Edmore is a Data Engineer at the Standard Bank of South Africa. His main areas of focus are Data Integration, Data Transformation and partly Data Visualisation.

Rennie Naidoo is an Associate Professor of Informatics at the University of Pretoria, South Africa. His research focuses on the interplay between information systems, people and organizations. His papers have appeared in the Journal of Strategic Information Systems, the European Journal of Information Systems, Information Technology & People, and The Information Society Journal.

Mehwish Nasim is a Computer Science Lecturer at the University of Western Australia. She works in the emerging field of social cybersecurity. Informed by sociological theories on human interaction, she develops mathematical models of social influence to combat the dissemination of misinformation across social media platforms. She also models operations in information environment in wargames.

Tabisa Ncubukezi is an academic at the Cape Peninsula University of Technology. She has done her Ph.D. in the same institution. Her primary research areas are communication networks, cloud computing, cyber security, and e-Learning.

Tatu Niskanen is a researcher with a background in cybersecurity and information systems science. In addition to research work, he has experience in software development. He is working for the Applied cybersecurity team at VTT Technical Research Centre of Finland and finishing up his master's degree at the university of Jyväskylä.

Daniel Nistor is a Public Affairs specialist with more than a decade of experience in media operations for the Romanian Armed Forces and Ministry of Defence. He is currently a PhD student at the National Defence University, conducting a research on consolidating a society's security culture through better institutional strategic communication.

Nombeko Ntingi holds a BSc (Computer Science) from the University of Transkei, Master's degree in IT from Carnegie Mellon University, and a BCom (Financial Management) from the University of South Africa. She holds certificates in SAP ABAP, Cyber Security from UJ, CPUT and F'SAT; Managing Risk Harvard Online. She is currently pursuing PhD studies at the University of Johannesburg.

Victor Kwarteng Owusu is a Ph.D. candidate at the Faculty of Management and Economics, Tomas Bata University in Zlin, Czech Republic. He has two publications to his credit. He has also attended conferences across Europe and the United Kingdom including British Academy of management (BAM), hosted by Manchester business school. His main research area is organizational performance.

Jani Päijänen works as a Project Manager at the Institute of Information Technology of JAMK University of Applied Sciences. His has industry background in delivering consultancy for clients in Project Management, Information Technology, and Software Development.

Tehi Palletvuori is Security Management master's degree student at Laurea University of Applied Sciences. Palletvuori BBA (Finance), MBA (Entrepreneurship and Business Competence) has over 20 years of working experience in Banking, Corporate Treasury and IT. Since March 2020 she has worked at Finnish National Emergency Supply Agency and is responsible for financial sector preparedness.

Nikolaos Papakonstantinou, D.Sc.(Tech.), Docent, is an Electrical and Computer Engineer interested in system resilience and cybersecurity. He is the Research Team Leader of the Applied cybersecurity team at the VTT Technical Research Centre of Finland.

Konstantina Papachristopoulou is a senior Research ICT Project Manager for EU-funded projects in Eight Bells Ltd. She has a diverse background and R&D experience obtained from various initiatives across European and national programs. She has successfully delivered more than 25 multidisciplinary research projects, focusing on areas such as cybersecurity, IoT & CPSoS, health, and sustainable manufacturing.

Ada Peter is a Professor, Digital Infrastructure, and International Security. She produces comprehensive, policy-focused, nontechnical cybersecurity documents that emphasize and teach senior principal state officials how to make informed cyber policy planning decisions. Her current research focuses on how U.S. national cyber strategies undermine critical infrastructures' protection against previous and potential cyber weapons.

Lucas Potter earned his B.S. in Biomedical Engineering in 2015 from VCU and a PhD in Biomedical Engineering ODU. He has researched human factors for three federal organizations, biomedical engineering for three private companies and is an active contributor to the field of Biocybersecurity.

Dr. Jouni Pöyhönen, Col (ret.) is a postdoctoral researcher of cyber security programs in University of Jyväskylä. He received a Ph.D. in Information Technology from University of Jyväskylä. He has over 30 years' experience as developer and leader of C4ISR Systems in Finnish Air Forces. He has more than twenty cyber security research papers or articles.

Jyri Rajamäki, Principal Lecturer in Information Technology at Laurea University of Applied Sciences and Adjunct Professor of Critical Infrastructure Protection and Cyber Security at the University of Jyväskylä, Finland, holds D.Sc. degrees in electrical and communications engineering from the Helsinki University of Technology, and a PhD in mathematical information technology from the University of Jyväskylä.

Trishana Ramluckan is the research manager at Educor Holdings, an honorary research fellow for Law at the University of KwaZulu-Natal, a curator of the EU:CD's Cyber Diplomacy Atlas, a member of the IFIP working group on ICT Uses in Peace and War, and was listed in the Top 50 Women in Cybersecurity in Africa (2020).

Huw O.L. Read is Charles A. Dana Professor of Digital Forensics and the Director of the Centre of Cybersecurity and Forensics Education and Research (CyFER) at Norwich University, Vermont, USA. For over 15 years, he has taught in several countries, authored over 20 peer-reviewed publications in the field and has been awarded numerous competitive grants.

Aunshul Rege is an Associate Professor and Director of the Cybersecurity in Application, Research, and Education (CARE) Lab at Temple University. Her research focuses on the human and behavioral aspects of cyberattacks, surveillance, privacy, social engineering, and cybersecurity education, and has been funded by various National Science Foundation grants.

Mari Ristolainen is a Researcher at the Finnish Defence Research Agency. She has studied psychology at the Moscow State University and she earned a doctorate in Russian Language and Cultural Studies from the University of Joensuu in 2008. She has been conducting postdoctoral research in the field of Russian and Border Studies in several Academy of Finland- and EU-funded projects at the University of Eastern Finland and at the University of Tromso, Norway. Her current research interests include cyber warfare as a phenomenon, Russian digital sovereignty, and the governance of cyber/information space.

Harri Ruoslahti, PhD, works as a Principal lecturer at the Resilience Lab of Laurea University of Applied Sciences.

Helvi Salminen has worked in information security since June 1990 in a high security industrial environment. Before her security career she has 12 years of experience in systems development. She has 30+ years experience in cryptographic systems management and various security frameworks. She was awarded as CISO of the year in Finland 2014.

Nkata Sekonya is a postgraduate student in computer science at the University of Johannesburg, SA. He received his BSc in Information Technology from the University of Johannesburg in 2021. His main research interests are digital forensics, critical infrastructure protection, and the internet of things.

Radu Serrano is PhD student at the Ragnar Nurkse Department of Innovation and Governance of TalTech, Estonia. He also develops and manages cybersecurity, smart governance, e-democracy and technology projects at the e-Governance Academy. His main research areas include cybersecurity, internet voting and business administration.

Jussi Simola: post-doctoral researcher and has a PhD degree in cyber security from the University of Jyväskylä. Expertise includes decision-support technologies, situational awareness systems, information security, and continuity management. His current research work is focused on the development of maritime cyber security. Has been involved in developing a common early warning system for the EC.

Ms Thenjiwe Sithole is a PhD student at the University of Johannesburg. She holds a Masters in Information Technology (Information Systems) from the University of Pretoria and a Master of Engineering Sciences in Electronics (Telecommunications) from the University of Stellenbosch. She also has Certificate in Cyber Security from the University of Johannesburg.

Siphesihle Sithungu is from Johannesburg, South Africa and holds an MSc in Computer Science (University of Johannesburg). He is a lecturer at the University of Johannesburg, and his research interests are artificial intelligence and critical information infrastructure protection. Mr Sithungu is a technical committee member for the International Conference on Computational Intelligence and Intelligent Systems.

Adam Spanier is currently pursuing a PhD in IT Computing Systems from the University of Nebraska at Omaha. He has a BS in Computer Science and is pursuing his MS in Cybersecurity concurrently with his PhD. His research interests include intermediate representation analysis in cybersecurity, computational trust, and cybersecurity gamification.

Iain Sutherland is Professor of Digital Forensics and Head of Research at Noroff University College in Kristiansand, Norway, a university he helped to establish in 2012. A recognised expert in computer forensics, he has supervised several PhD students and provided evidence as an expert witness. He has authored articles ranging from forensics practice and procedure to network security.

Pardis Moslemzadeh Tehrani is the Associate Head of the Law School, at the Faculty of Business, Law and Tourism, University of Sunderland. Before joining the Sunderland University, she was a Senior Lecturer and Visiting Associate Professor at the Faculty of Law, University of Malaya. Her research spans the area of information technology law, international law and legal research methodology. She is the author and co-author of more than 70 research papers, books, book chapters and conference papers. She has served on many conferences, seminars, forums and workshop programme committees and delivered talks at various panels, NGOs, research centres and international bodies such as ICRC and other Human Rights organisations.

Ranko Trišić. I started my Master's degree, in Games Studies and Engineering, in 2020, in Klagenfurt, Austria. My research and thesis are about hacking and cheating in game industries. As a part of my thesis, I had to create game according to topic. So, I created board game under the name "*HACKER'S CYBER ATTACK".*

Uakomba Uhongora is a PhD candidate at the University of South Australia. Her PhD is on space systems cyber security with a focus on the integration of SDN and space systems. She is a lecturer at the Namibia University of Science and Technology. Her research interests are cyber security, deep learning, space systems and SDN.

Brett van Niekerk is a senior lecturer at the Durban University of Technology, chairs the International Federation of Information Processing Working Group on ICT in Peace and War, and is co-Editor-in-Chief of the *International Journal of Cyber Warfare and Terrorism*. He is CISM certified, with over 50 publications and 20 presentations.

Basie von Solms is a Research Professor in the Academy for Computer Science and Software Engineering at the University of Johannesburg. He specialises in Cyber Security. He is a Past President of the International Federation for Information Processing (IFIP). Professor Rossouw von Solms is an emeritus professor from the Nelson Mandela University. Rossouw has supervised 65 students successfully to a Master's and Doctoral degrees in the fields of information security and IT governance. He has published and presented in excess of 200 peer-reviewed research papers.

Richard L. Wilson is a Professor in Philosophy at Towson University in Towson, MD. Professor Wilson Teaches Applied Ethics in the Philosophy and Computer and Information Sciences departments at Towson while also serving as Senior Research Scholar in the Hoffberger Center for Professional Ethics at the University of Baltimore. Professor Wilson's interest's are

directed towards applying phenomenology to issues with emerging and innovative technologies.

Allison Wylde holds posts with Glasgow Caledonian University (London) as a Senior Lecturer, and with the Cardiff University Centre for Cyber Security Research: research includes- artificial intelligence (AI) intelligent virtual assistants (IVAs); trust and zero trust; and, work on cyber security norms and the Global Digital Compact with the United Nations Internet Governance Forum.

Saliha Yilmaz is currently pursuing a PhD in Cybersecurity at Kingston University London. Her research focuses on detecting privacy risk-prone mobile applications. Her previous research and studies focused on predictive policing and personal data legislation in government institutions. Cybersecurity, data privacy, threat intelligence, crime science, and data-driven policing are among her research interests.

A Survey on National Cyber Emergency Plans

Konstantinos Adamos[1,2], Ioannis Filippopoulos[1], George Stergiopoulos[1] and Dimitris Gritzalis[2]

[1] Dept. of Information and Communication Systems Engineering, University of the Aegean, Samos, Greece

[2] Dept. of Informatics, Athens University of Economics and Business, Athens, Greece

kadamos@aegean.gr
icsdm320027@icsd.aegean.gr
g.stergiopoulos@aegean.gr
dgrit@aueb.gr

Abstract: Operators of Essential services (OESs) and Critical infrastructures (CIs), whether private companies or public organizations are going through a digital transformation to pace with the evolution of technology and to bring better services to customers and countries' citizens. Operational Technology (OT) systems like Supervisory Control and Data Acquisition (SCADA) and Distributed Control Systems (DCS) used to control and monitor functions in such infrastructures are converging with Information Technology (IT) environments. This convergence has exposed infrastructures to new cyber risks. For this reason, EU Member States have been trying to build resilience against cyber-attacks to ensure the stable operation of their states. Several countries have established cybersecurity incident response procedures as well as steps or phases of response before, during, and after a cyber incident. The sum of these procedures and guidelines constitutes their national cyber emergency plans (NCEPs). Still, these NCEPs differ widely in their approaches. These differences manifest as both managerial, governmental, legal, and technical, creating a complex environment worldwide. In this paper, we gather four major NCEPs worldwide to analyze and compare them with prominent standards and industry guidelines in cybersecurity, like the ISO 27001 and NIST 800 series. We investigate NCEP approaches to building cyber resilience based on their response models, their involved entities, the cooperation between agencies and other countries, and their risk-based categorization for cyber incidents. We elaborate on their differences, potential issues and divergences and argue whether these plans can be combined to bridge potential weaknesses. We selected and surveyed four (4) cyber emergency plans from four (4) countries that are frequent targets of cyber-attacks and have long experience in managing and responding to cyber incidents.

Keywords: Cybersecurity**,** Critical Infrastructure, Incident Response, National Cyber Emergency Plan, Operational Technology, Information Technology

1. Introduction

Cyber-attacks against national CIs are on the rise for the past decade (Lohrmann, 2022) (Trend-Micro, 2022) and cyber warfare has intensified (Claridge, 2022) (Pearson & Bing, 2022). Countries try to cope with this trend by build resilience to cyber-attacks to protect their most vital assets. This procedure is formalized by establishing national cybersecurity strategies and national cyber emergency plans. A National Cyber Emergency Plan (NCEP) is a set of guidelines, methodologies, and practices (usually following specific standard guidelines) that are formally applied by countries to respond to significant cyber incidents. In this paper, we survey four (4) prominent (in terms of cybersecurity posture) NCEPs and analyze their approaches concerning incident response procedures, their cyber incidents categorization, and their roles and responsibilities in case of significant cyber incidents. Specifically, NCEPs from Germany, France, the UK, and the US are analyzed and cross-compared for the first time in literature, along with comparative data for all different NCEPs. This work provides an insight into differences and potential weaknesses of major NCEPs and proposes combinatory implementations to address documented weaknesses or risks. These data can be used by other countries or organizations to compare and update their NCEPs to remove potential weaknesses.

In Section 2 we briefly present key points in four national cybersecurity strategies, NCEPs, and their objectives and present related work in the area. In Section 3 we present each country's CI sectors and in Section 4 each country's way of categorizing cyber incidents. In Section 5 we describe the roles and responsibilities within NCEPS while in Section 6 we present the national cyber incident response procedures. Finally, in Section 7, we depict comparative data of the surveyed countries and summarize best practices, convergences, and divergences, and in Section we Discuss preliminary findings and conclude our analysis.

2. Related work

2.1 National strategies

Concerning National cybersecurity strategies and cyber emergency plans, Germany, France, the UK, and the US are four major countries that issued theoretical baselines to support cybersecurity strategies to set out their objectives. Still, only two (2) of them have formally established distinct NCEPs, in the sense of a document that precisely describes the processes in case of a significant cyber incident. This section briefly mentions the four (4) countries' national cybersecurity strategies and cyber emergency plans.

Since 2011, Germany has adopted a comprehensive cybersecurity strategy, which is frequently updated with the aim of the uninterrupted and secure operation of its CIs. The latest version of the German Cybersecurity Strategy was published in 2021 (BMI, 2021). The strategy covers the next five years and includes guidelines for the protection of organizations and a faster response to cyber incidents. In addition, it clearly describes the national goal-setting at a strategic level and focuses, among other things, on the promotion of cooperation (BMI, 2021). The country has not yet issued a distinct NCEP.

The latest French cybersecurity strategy was published in 2015 and includes five (5) strategic objectives to keep the nation cyber-secure (SSI, 2015). France has established an NCEP since 2004, known as the "Piranet Plan", which is top-secret and is only tested during exercises (Brangetto, 2015).

The UK's revised 2022 Cybersecurity Policy points to a future where the UK is even more resilient to cyber-attacks. The strategy sets out five (5) pillars to keep the UK cyber-secure (HM Government, 2022). The UK has not yet issued a distinct NCEP.

The US gives special importance to the defense of its national interests, a fact that is also reflected in the cybersecurity field, since it is one of the few countries that, in addition to its national cybersecurity strategy, has also issued a distinctive NCEP (i.e., the National Cyber Incident Response Plan - NCIPR).

The NCIRP follows five (5) basic principles to respond to incidents in the public and private sectors. (Department of Homeland Security, 2016).

2.2 Literature work

Early work on national processes concerning cyberattacks was published by Harrop and Matteson (Harrop, 2015). In their work survey potential cyber-attacks on critical national infrastructure in the United Kingdom and the USA, and discuss security measures that were drafted by these countries as mitigations controls against such threats. Albeit, important, this work does not focus on analyzing entire NCEPs, which were did not exist in their current form when this work was published. Also, authors do not critically analyze nor compare implemented approaches and lessons learned between countries, but mostly emphasize on future directions and trends in these two countries.

Work from Daricili and Cerik (Daricili, 2021) explores the relationship between the concept of national security and cyber threats, along with a theoretical analysis of cybersecurity based on the national perspective of the US. The study includes definitions of related concepts from the Turkish national cyber-defense program.

A Comparative Analysis of twenty National Cyber Security Strategies was published in 2015 (, 2015). This work aimed to use common metrics as reference to compare the cybersecurity readiness of multiple nations. Still, this work was published before the implementation of full NCEPs by major nations and mostly compares definitions, responsible authorities and types of threat characterization by each country in the study. It does not advance to more practical concepts, such as assessment process comparison, technical comparisons and implementation procedures of emergency plans (Shafqat, 2016).

3. National critical infrastructure sectors

Germany created a distinct national strategy for Cis, known as the KRITIS plan (BMI, 2021). In addition, the country created a more specialized strategy (i.e., the National Strategy for Critical Information Infrastructure Protection - CIIP Strategy) which includes structures related to IT. The competent bodies for the implementation of the aforementioned strategies are the Federal Office of Civil Protection and Disaster Assistance (BBK) and the Federal Office for Information Security (BSI) (BSI, n.d.). Germany has identified ten (10) key critical infrastructure sectors (see Table 1).

2013 marked a turning point for France's national cybersecurity strategy, as the new regulatory framework known as the CIs Information Protection (CIIP) Law was implemented. Due to the rapid increase in the number and sophistication of cyber-attacks, the country defined through the CIIP Law the critical operators, but also the obligations that these operators must fulfill. France recognized over 200 public and private critical operators from 12 CI sectors (See Table 1). (ANSSI, n.d.)

The UK has defined 13 national CI sectors (see Table 1). The responsible government body for the protection of the national CI of the UK is the Center for the Protection of National Infrastructure (CPNI). (CPNI, 2021)

The US has defined 16 CI sectors (see Table 1), the protection of which is vital for its national interests. For each sector, there is a self-governing council known as the Sector Coordinating Council (SCC) (CISA, n.d.), whose members include owners and operators of the infrastructure sectors. SCCs participate in various forums for members of all sectors to interact with each other and pursue a common policy and strategy in their efforts to ensure their resilience, including in cyberspace. (CISA, n.d.)

Table 1: Identified CI Sectors per country

Country / CI Sectors	Germany (10)	France (12)	UK (13)	USA (16)
Chemicals			X	X
Dams				X
Emergency Services			X	X
Energy	X	X	X	X
Finance	X	X	X	X
Food - Agriculture	X	X	X	X
Government - State - Administration	X	X	X	X
Health	X	X	X	X
Industry - Manufacturing - Commercial facilities		X		X
Information Technology – Communications - Broadcasting	X	X	X	X
Justice		X		
Media and Culture	X			
Military - Defence		X	X	X
Waste	X			X
Nuclear			X	X
Space		X	X	
Transport	X	X	X	X
Water	X	X	X	X

4. Cyber-incident categorization

Germany follows the BSI standard 100-4, issued in November 2008 and still in effect. In January 2021, the draft of the new standard 200-4 (BSI, 2021) was published in its 1st edition (a 2nd draft version is expected). Six (6) escalation levels are classified (see Table 2). From disaster level 6 large-scale damaging events that are not restricted to the affected entity alone to normal operation level 1 minor events. (BSI, 2009)

France follows a variant of the corresponding American scheme. Specifically, it classifies incidents into six (6) levels (level 1 is divided into 1A and 1B) (see Table 2). France takes a more "aggressive" stance in level 5 cyber incidents, which significantly differentiates it from other countries. Specifically, *these incidents are equated with "armed conflict" cases, which according to article 51 of the Charter of the United Nations, include loss of life or large-scale physical disasters*. (SGDSN, 2018)

The National Cyber Security Center (NCSC) updated UK's incident categorization system in April 2018 by increasing the categories from three (3) to six (6) (see Table 2). In this system, all incidents are included regardless of the size and role of the affected entity. A gradual involvement of agencies is noticed, during the escalation of an incident. Nonetheless, in almost all cases the response is given by the affected entity. The

exception lies in level 1 incidents, i.e., cases of national emergencies, where due to the very high severity, the response is undertaken by the government at a strategic level and the NCSC at a tactical level. (NCSC, 2018)

It is noteworthy that *level 1 and 2 incidents, combining likelihood and impact, are considered by the UK more dangerous than threats of nuclear attacks and as dangerous as international terrorism, large-scale accidents, physical disasters, and international military crises* (HM Government, 2015).

The US NCEP adopts a scheme of escalating the severity of cyber incidents. This scheme ensures that all services and agencies have a common picture of the severity of an incident to respond accordingly. According to the scheme, six (6) incident severity levels are defined, marked with different coloring (see Table 2). From level 0 (baseline - white color) which is an unsubstantiated or inconsequential event to level 5 (emergency - black color) which poses an imminent threat to the provision of wide-scale CI services, national government stability, or to the lives of U.S. persons. (The White House, n.d.) The Cybersecurity and Infrastructure Security Agency (CISA) created a national incident scoring system (National Cyber Incident Scoring System - NCISS), to rate the severity of an incident (CISA, n.d.).

Table 2: Cybersecurity Incident Severity Scale

Germany		France		UK		US	
#	Severity level	#	Severity level	#	Severity level	#	Severity level
6	Disaster	5	Situation of Extreme Emergency	1	National cyber emergency	5	Emergency
5	Crisis	4	Major Crisis	2	Highly significant incident	4	Severe
4	Emergency	3	Crisis	3	Significant incident	3	High
3	Early warning	2	Serious Incident	4	Substantial incident	2	Medium
2	Fault alarms	1B	Incident	5	Moderate incident	1	Low
		1A	Event				
1	Normal operation	0	Minor Event	6	Localised incident	0	Baseline

5. Roles and responsibilities

Germany follows a h*olistic approach to incident response, involving multiple agencies, entities, institutions, and even individuals*, each responsible for a part of the response process. Given that a significant incident may affect several critical sectors, the individual response per sector by different agencies is considered insufficient. Therefore, the country allocates responsibilities to the affected entities, federal agencies, and the private sector, encouraging cordial cooperation between them. To support this structure, Germany implemented a multitude of government services and agencies. It is indicative that it has staffed 57 services at the European level and 17 within the NATO framework. It has also established *67 agencies at the federal level, 34 at the state level, and 5 at the local level* (BMI, 2021) (Beigel & Herpig, 2021), the highest number in all study material.

The French National Agency for Cyber Security (ANSSI) is the official national agency for responding to cyber incidents. ANSSI belongs to the General Secretariat of the Ministry of Defense and National Security (SGDSN), an inter-ministerial body under the Prime Minister that undertakes, among other tasks, the management of national crises at a strategic level. Also, the Ministry of Defense has an active role in the country's incident response. It is noteworthy that the type of incidents reported by critical operators depends on the CI they manage. In particular, the ministry to which the respective sector of CI falls is the one that determines the type of incidents that must be reported to ANSSI. Also, this ministry assumes a coordinating role during the occurrence of incidents and defines the guidelines, objectives, contact points, vulnerabilities, and risks faced by each of its sectors, at the national level (SGDSN, n.d.).

Responsible for the UK's protection of data, networks, and CI information systems is the NCSC, which is also the cyber incident response national agency. CPNI and NCSC work closely together in the broader framework of CI security. The NCSC has an advisory role and essentially focuses on limiting the spread of a cyber incident, within the boundaries of the affected entity, to avoid a national crisis. Within the entity, if it belongs to the private sector the entity itself is solely responsible, while if it belongs to the public sector the responsibilities are distributed both to the entity and to the ministry to which it belongs. The country defends the view that "an entity that has fallen victim to a cyber-attack is also responsible for the consequences". Therefore, the NCSC is

limited to advise and analysis and only in exceptional cases deploys a response team and provides on-site support (HM Government, 2022).

In the US, the federal government has a leading role in managing the response to significant cyber incidents on entities (private or public). US NCIPR focuses on 4 lines of effort in which different federal agencies are involved. The lead federal agencies for the threat response line of effort are the Federal Bureau of Investigation (FBI) and the National Cyber Investigative Joint Task Force (NCIJTF) and they belong to the Department of Justice (DOJ). For asset response, the National Cybersecurity and Communications Integration Center (NCCIC) takes over and it belongs to the Department of Homeland Security (DHS), which also includes the US-CERT and the ICS-CERT. For intelligence support, the Cyber Threat Intelligence Integration Center (CTIIC) is the lead agency and it belongs to the Office of the Director of National Intelligence (ODNI). Finally, for the affected entity response effort, if a significant cyber incident affects a federal agency, then the agency has the lead responsibility for the response, while if a private entity is affected, then the entity has the lead responsibility for the response and the corresponding sector-specific agencies will assist the Federal Government to understand the potential impact on CIs.

6. Incident response procedure

Germany's response against significant cyber incidents is based on the resilience of the affected entities and the communication between the involved actors. Appropriate preparation and immediacy in information are the key points that will initially ensure the limitation and finally, the elimination of the crisis. For this reason, it implements a national IT Early Warning System (EWS) (Kossakowski, et al., 2006).].

The incident response steps include the implementation of an incident response plan within the entity, reporting of the incident to BSI, notification of the affected individuals, cyber incident assessment by CERT-Bund's IT-LZ, notification of other entities (potential targets) through Cyber-AZ, supporting the affected entity through CERT-Bund and MIRT, activation of the national crisis center IT-KRZ and finally requesting for external support, if needed. (BMI, 2005) (BSI, 2009) (BSI, n.d.) (BSI, n.d.)

France gives great importance to the exchange of information during cyber incidents. Critical operators send a notification form in the event of an incident, while ANSSI responds by providing suggestions or on-site support. In addition, affected entities may voluntarily exchange information with other critical operators. All information exchanges in which ANSSI participates are classified as confidential and are handled accordingly. Also, in addition to the response against the attacker, the country assumes a more active involvement within the affected entity. This implies that in the event of a significant incident if the country deems it necessary by a decision of the Prime Minister it acquires additional jurisdiction to undertake any action. It could also take control of the information systems of the affected entity, regardless of whether it belongs to the public or private sector. (ANSSI, n.d.) (Brangetto, 2015).

UK's NCSC suggests to entities the preparation of an indicative plan when a significant cyber incident takes place. To respond successfully, the entity must follow the steps below:

Step 1: Understand the type and severity of the incident.
Step 2: Kick-off Response – Reporting.
Step 3: Incident Management.
Step 4: Escalation and decision making.
Step 5: Review and close down - lessons learned. (NCSC, n.d.)

In the US, public or private entities must have cyber incident response plans in place to protect their assets and operations. In the case of a cyber incident, the affected entity must initially make every possible effort to limit the effects and secure its assets. When an incident endangers the entity's CIA, resulting in significant damage to business functions, the entity must submit a report to the corresponding federal agency, by providing basic information about the incident. The US legislation does not make reporting mandatory for the private sector, while government agencies are required to report within a specific time frame. In the event of a significant cyber incident, the Cyber Unified Coordination Group Cyber (UGC) undertakes the operational coordination of the incident response. (Department of Homeland Security, 2016)

7. Best practices – Convergences and divergences

The analysis of the NCEPs revealed significant differentiations between countries. Even within the EU, where European legislation is common to the member states, there is a certain difference in opinions and practices on important issues. Of course, there are also areas in which most member states follow the same line, for example in the selection of CIs, where no notable discrepancies are found.

7.1 Comparison of incident handling models

The models applied at a national level are mainly divided into centralized and decentralized. A centralized model applies in EU countries (i.e., France and Germany). When significant incidents occur, the competent agencies take full control of specific actions, having a correspondingly increased share of responsibility. On the other hand, in the US and the UK, the decentralized model applies. This model provides a more flexible policy with government agencies playing a more consultative role and responsibilities being shared between the involved entities.

It is worth noting that the decentralized model of the UK and the US follow a different way of diffusing the responsibility that arises from managing an incident. The risk of incident handling is often shifted to private companies or agencies that undertake the handling of an incident (perhaps entirely), without the competent agencies imposing or undertaking an active or interventional role (HM Government, 2022) (Department of Homeland Security, 2016).

In contrast, the centralized models of France and Germany take an active role in incident handling and often impose procedures and organizational security measures on the actors involved, even under the threat of fines in case they do not fully comply.

7.2 Distinct national cyber emergency plans

An important issue that was identified is that most countries have not issued a distinct NCEP to handle cybersecurity incidents and simply include instructions attached to their national strategy or in separate documents. For example, Germany has a plan for national crises (not for cyberspace) and supports many IT standards through the BSI, but no concise and comprehensive text describes the sequence of actions when significant cybersecurity incidents occur.

On the other hand, the US issued a comprehensive NCEP that clearly and accurately describes the actions of the involved entities. France also has a distinct plan, which is top-secret, while the UK is based on a set of directives, similar to Germany's.

7.3 Involved entities

Differentiations are also ascertained in the number of organizations involved when a large-scale cyber-attack takes place. For example, Germany as a federal state, involves a large number of agencies, making it difficult to determine jurisdictions and the level of involvement between these agencies. This tactic is not only a waste of resources, but it makes the system rigid, with little adaptability to the ever-changing conditions of cyberspace. In addition, the information exchange gets more difficult, since the number of recipients increases and by extension, the possibility of incorrect information or omission of an important recipient also increases.

On the contrary, France and the UK involve significantly fewer agencies. France has fully staffed the corresponding government agencies in recent years. These different approaches to the organization and structure of the country response mechanism are estimated to be mainly due to the structure of the country itself (e.g., federation), the geopolitical environment (e.g., neighboring threats that increase the possibility of cyber-attacks), and the large fluctuations of available funds and personnel.

Regarding the government and private sector relationship, some of the aforementioned countries use control mechanisms to enforce their policies, while others rely on the cooperation and initiative of the private sector. For example, Germany (less) and France (more) take a strict posture towards private companies that manage CIs. According to the law, specific or minimal measures must be taken by private companies to protect their sensitive information systems. At the same time, government agencies conduct regular audits on the implementation of these measures.

Especially France has legislated the direct involvement of state bodies in the internal systems of companies, in case of national interest reasons (Brangetto, 2015) (Cymutta, 2020). In contrast, the US and the UK apply looser legislation and invest in mutual trust and voluntary implementation of the proposed directives, hoping that the private sector will take the necessary protective measures on its own.

7.4 Involvement of military services

Another important factor worth mentioning is the involvement of the Ministry of Defense and in general the military sector of the states in dealing with cyber-attacks. For example, France actively involves the military in incidents categorized at the highest level of the country's classification scale, thus seeking to give a different dimension to incident response methods, both substantively (increased military technological capabilities) and semantically (escalating the reaction and intimidating the attacker). In case a cyber-attack was instigated by another country, France favors a more aggressive than defensive response, even with military involvement. Regarding the other countries, the military authorities are not actively involved, except in Germany, where they participate in cyber exercises for coordination and information purposes.

7.5 Incident classification

Concerning cyber incident classification, all countries distinguish six (6) basic levels of severity, with minor differences in their designation. In practice, however, no incident level is identical, nor does it escalate in the same way. Trying to fit it an incident into similar levels in different national classifications proved to be a difficult task that requires additional resources and does not necessarily capture the essence of each country's response similarities. Therefore, there is a high probability that it will lead to conclusions. On the other hand, incidents are classified mainly for practical reasons, i.e., that all involved entities "speak" the same language and that there are automated procedures and pre-defined response steps, to be able to extract statistical data for use in future threats (ENISA, 2010). To this end, we built a comparative table of the incident classification of the four (4) government entities (see Table 3) that allows for multiple cross-comparison of a specific incident against multiple national levels.

Table 3: Comparative case classification data of the surveyed countries

Case \ Country	Germany	France	UK	USA
No. of incident severity levels	6	6	6	6
Incident severity levels	1. Disaster 2. Crisis 3. Emergency 4. Early warning 5. False alarm 6. Normal functions	1. Extreme emergency 2. Big Crisis 3. Crisis 4. Significant incident 5. Event 6. Small event	1. National cyber emergency 2. Highly significant incident 3. Significant incident 4. Substantial incident 5. Moderate incident 6. Localised incident	1. Emergency 2. Severe 3. High 4. Medium 5. Low 6. Baseline
Important classification criteria	- Impact on critical assets or basic functions of the entity. - Maximum tolerable time of loss. - Expanding impact outside the entity.	- Impact on critical assets or basic functions of the entity. - Intention of the attacker. - Expanding impact outside the entity.	- Impact on critical assets or basic functions of the entity. - Expanding impact outside the entity.	- Impact on critical assets or basic functions of the entity. - Observed activity and intention of the attacker. - Expanding impact outside the entity. - Loss of human lives. - Country destabilization.

Country / Case	Germany	France	UK	USA
External support	Incident levels 1 and 2.	Incident levels 1, 2, and 3 (At level 1 the national authority acquires full access).	Incident levels 1 and 2.	Incident levels 1 and 2.
Simple Guidance	Incident levels 3, 4, and 5.	Incident levels 4, 5, and 6.	Incident levels 3, 4, 5, and 6.	Incident levels 3, 4, 5, and 6.

Preliminary results from indicative examples show that, across all countries, the highest two (2) levels (highest three (3) for France) are those that usually require external active support from the respective national body and concern situations with an impact at the national level, while the rest are most often eliminated by the affected entity with simple guidance of the national agency or other federal agency. Classification criteria remain largely similar across all NCEPs, although the quantitative scales differ, which is expected given the different size and focus on OES in these countries. **Error! Not a valid bookmark self-reference.** compares the most important criteria of the surveyed countries.

Table 4: Most important comparative criteria of the surveyed countries

Country / Criteria	Germany	France	UK	USA
Number of CIs	10	12	13	16
Model	Centralized		Decentralized	
Distinct national plan	Yes	Yes (confidential)	No	No
National response agency	BSI	ANSSI	NCSC	NCCIC, CTIIC
Role in significant incidents	Coordinator	Coordinator - Managerial	Consultant - Coordinator	
Structure - organizing	Rigid		Flexible	
Military involvement	Yes		No	
On-site support	Yes			
Measures were taken by entities	Mandatory		Optional	

8. Discussion and Conclusions

Each NCEP reflects the countries on tendencies and public sector organization. Countries, such as the UK and the US focus more on an audit-based approach while leaving enough room to organizations to build their own cyber security posture. Still, major fines and legal consequences are due in case organizations prove to be unready to major attacks, especially when these attacks impact their national citizens.

Contrary to this, Germany and France follow a more “public-sector based” approach, where the public sector and the government have an active role in managing and micro-managing the implementation of cyber-security processes and measures.

Different approaches are also found regarding the willingness of the examined states to cooperate or even request support from other states and international organizations when dealing with significant cyber incidents. For example, the UK is cautious about the possibility of cooperating with another country against a known cyberthreat, even if it is ascertained that the source of the cyber-attack is in this country (Collier, 2017). Although the UK is actively involved in international efforts to tackle cybercrime, it appears to distrust other countries to the extent of allowing them to participate in addressing a national crisis on its soil, especially after Brexit. On the contrary, countries such as Germany and France base their defense on partnerships and seek common responses both at the national and European levels.

The current perspective shows that both management models, whether government-centered (e.g. Germany) or industry-centered (UK) suffer from major issues concerning their viability against different types of cyber threats. For example, the federated approach of Germany is too slow to react in fast-paced cyber-attacks that hit multiple organizations at the same time. On the other hand, the UK’s approach to allowing organizations to tackle their own threats as they see fit, leaves room for error and negligence, especially in SMEs, while at the same time creating fertile ground for information isolation and lack of information flow in terms of coordination

and clarity. The UK and US have a mandate to inform the state of potential cyber-incidents, but history has shown that often, companies are reluctant to share information and disclose their cyber-breaches weeks or even months after they have taken place, which leaves enough time for attackers to advance and capitalize on their spoils.

It seems that a more hybrid approach, where the government plays an active role in enforcing and auditing organizations, especially OES and Cis, while at the same time allowing their leadership the responsibility required to act in time and adjust security measures and procedures to each one's way-of-work, creates the best intersection of both models.

Acknowledgements

This work has been supported in part by a Research Grand on Cybersecurity offered by the Hellenic Ministry of Digital Transformation to the Athens University of Economics and Business, Greece (2022-24).

References

ANSSI, n.d. CIIP IN FRANCE - FAQ. [Online]
Available at: https://www.ssi.gouv.fr/en/cybersecurity-in-france/ciip-in-france/faq/
[Accessed 09 2022].

Beigel, R. & Herpig, S., 2021. Germany's Cybersecurity Architecture. [Online]
Available at: https://www.stiftung-nv.de/sites/default/files/eng_impulse-germanys_cybersecurity_architecture_translation_of_the_6th_german_edition_0.pdf
[Accessed 09 2022].

BMI, 2005. National Plan For Information Infrastructure Protection. [Online]
Available at: https://www.qcert.org/sites/default/files/public/documents/GER-PL-National%20Plan%20For%20Information%20Infrastructure%20Protection-Eng-2005.pdf
[Accessed 09 2022].

BMI, 2021. Cybersicherheitsstrategie für Deutschland 2021. [Online]
Available at: https://www.bmi.bund.de/SharedDocs/downloads/DE/veroeffentlichungen/2021/09/cybersicherheitsstrategie-2021.pdf;jsessionid=D8CB089E599784260FAD110D0DD36D39.1_cid373?__blob=publicationFile&v=1
[Accessed 09 2022].

Brangetto, P., 2015. National Cybersecurity Organisation: France. [Online]
Available at: https://ccdcoe.org/uploads/2018/10/CS_organisation_FRANCE_032015_0.pdf
[Accessed 09 2022].

BSI, 2009. BSI-Standard 100-4. [Online]
Available at: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/BSIStandards/standard_100-4_e_pdf.pdf?__blob=publicationFile&v=1
[Accessed 09 2022].

BSI, 2009. BSI-Standard 100-4. [Online]
Available at: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/BSIStandards/standard_100-4_e_pdf.pdf?__blob=publicationFile&v=1
[Accessed 09 2022].

BSI, 2021. BSI-Standard 200-4. [Online]
Available at: https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/BSI_Standards/standard_200_4_CD_2_0.pdf?__blob=publicationFile&v=5
[Accessed 09 2022].

BSI, n.d. Incident Response with CERT-Bund and MIRT. [Online]
Available at: https://www.bsi.bund.de/EN/Themen/Unternehmen-und-Organisationen/Cyber-Sicherheitslage/Reaktion/Vorfallunterstuetzung/MIRT/mirt.html
[Accessed 09 2022].

BSI, n.d. National IT Crisis Management. [Online]
Available at: https://www.bsi.bund.de/EN/Themen/KRITIS-und-regulierte-Unternehmen/Kritische-Infrastrukturen/Service-fuer-KRITIS-Betreiber/IT-Krisenreaktionszentrum/Nationales-IT-Krisenmanagement/nationales-it-krisenmanagement_node.html
[Accessed 10 2022].

BSI, n.d. What are Critical Infrastructures?. [Online]
Available at: https://www.bsi.bund.de/EN/Themen/KRITIS-und-regulierte-Unternehmen/Kritische-Infrastrukturen/Allgemeine-Infos-zu-KRITIS/allgemeine-infos-zu-kritis_node.html
[Accessed 09 2022].

CISA, n.d. CISA National Cyber Incident Scoring System. [Online]
Available at: https://www.cisa.gov/uscert/CISA-National-Cyber-Incident-Scoring-System
[Accessed 09 2022].
CISA, n.d. Critical Infrastructure Sectors. [Online]
Available at: https://www.cisa.gov/critical-infrastructure-sectors
[Accessed 09 2022].
CISA, n.d. Sector Coordinating Councils. [Online]
Available at: https://www.cisa.gov/sector-coordinating-councils
[Accessed 09 2022].
Claridge, D., 2022. Israel Bolsters Digital Defense amid Iran Cyber Threat. [Online]
Available at: https://www.geopoliticalmonitor.com/israel-bolsters-digital-defense-amid-iran-cyber-threat/
[Accessed 10 2022].
Collier, J., 2017. Strategies of cyber crisis management: Lessons from the approaches of Estonia and the United Kingdom. [Online]
Available at: https://www.politics.ox.ac.uk/sites/default/files/2022-03/strategies-of-cyber-crisis-management.pdf
[Accessed 09 2022].
CPNI, 2021. Critical National Infrastructure. [Online]
Available at: https://www.cpni.gov.uk/critical-national-infrastructure-0
[Accessed 09 2022].
Cymutta, S., 2020. CCDCOE - National Cybersecurity Organisation: GERMANY. [Online]
Available at: https://ccdcoe.org/uploads/2020/12/Country_Report_DEU.pdf
[Accessed 09 2022].
Department of Homeland Security, 2016. NATIONAL CYBER INCIDENT RESPONSE PLAN. [Online]
Available at: https://www.cisa.gov/uscert/sites/default/files/ncirp/National_Cyber_Incident_Response_Plan.pdf
[Accessed 09 2022].
ENISA, 2010. Good Practice Guide for Incident Management. [Online]
Available at: https://www.enisa.europa.eu/publications/good-practice-guide-for-incident-management
[Accessed 09 2022].
HM Government, 2015. National Security Strategy and Strategic Defence and Security Review 2015. [Online]
Available at:
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/478933/52309_Cm_9161_NSS_SD_Review_web_only.pdf
[Accessed 09 2022].
HM Government, 2022. National Cyber Strategy 2022. [Online]
Available at:
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1053023/national-cyber-strategy-amend.pdf
[Accessed 09 2022].
Kossakowski, K.-P., Sander, J., Grobauer, B. & Mehlau, J. I., 2006. CarmentiS A German Early Warning Information System-Challenges and Approaches-. [Online]
Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.123.6333&rep=rep1&type=pdf
[Accessed 09 2022].
Lohrmann, D., 2022. Cyber Attacks Against Critical Infrastructure Quietly Increase. [Online]
Available at: https://www.govtech.com/blogs/lohrmann-on-cybersecurity/cyber-attacks-against-critical-infrastructure-quietly-increase
[Accessed 10 2022].
NCSC, 2018. New Cyber Attack categorisation system to improve UK response to incidents. [Online]
Available at: https://www.ncsc.gov.uk/news/new-cyber-attack-categorisation-system-improve-uk-response-incidents
[Accessed 09 2022].
NCSC, n.d. Incident management. [Online]
Available at: https://www.ncsc.gov.uk/collection/incident-management/cyber-incident-response-processes
[Accessed 09 2022].
Pearson, J. & Bing, C., 2022. The cyber war between Ukraine and Russia: An overview. [Online]
Available at: https://www.reuters.com/world/europe/factbox-the-cyber-war-between-ukraine-russia-2022-05-10/
[Accessed 10 2022].
SGDSN, 2018. Revue Stratégique de Cyberdéfense. [Online]
Available at: http://www.sgdsn.gouv.fr/uploads/2018/02/20180206-np-revue-cyber-public-v3.3-publication.pdf
[Accessed 09 2022].
SGDSN, n.d. SGDSN IN ENGLISH. [Online]
Available at: http://www.sgdsn.gouv.fr/accueil/sgdsn-in-english/
[Accessed 09 2022].

SSI, 2015. French National Digital Security Strategy. [Online]
Available at: https://www.ssi.gouv.fr/uploads/2015/10/strategie_nationale_securite_numerique_en.pdf
[Accessed 09 2022].

The White House, n.d. Cyber Incident Severity Schema. [Online]
Available at:
https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Cyber%2BIncident%2BSeverity%2BSchema.pdf
[Accessed 09 2022].

Trend-Micro, 2022. The State of Industrial Cybersecurity, s.l.: Trend-Micro.

Wayne Harrop and Ashley Matteson. Cyber resilience: A review of critical national infrastructure and cyber- security protection measures applied in the uk and usa. In Current and Emerging Trends in Cyber Operations, pages 149‑166. Springer, 2015.

Daricili, A. Burak, and Soner Celik. "National Security 2.0: The Cyber Security of Critical Infrastructure." PERCEPTIONS: Journal of International Affairs 26.2 (2021): 259-276.

Shafqat, Narmeen, and Ashraf Masood. "Comparative analysis of various national cyber security strategies." International Journal of Computer Science and Information Security 14.1 (2016): 129-136.

Digital Forensic in a Virtual World: A Case of Metaverse and VR

Tayba Al Ali, Sara Alfulaiti, Manal Abuzour, Sheikha Almaqahami and Richard Ikuesan
Computing and Applied Technology Department, College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates

Tiiba-m7md@hotmail.com
sara.alfulaiti@hotmail.com
Manal_mohammed1970@outlook.com
sheikha113k@gmail.com
richard.ikuesan@zu.ac.ae

Abstract: Metaverse is a virtual space where users can interact with each other. It is a combination of virtual reality, augmented reality, and mixed reality. This evolving technology can offer many exciting opportunities that can be used for individuals and businesses. Although this technology has many advantages, people are misusing it for their benefit. Many cyberattacks are occurring in the metaverse world because it has various vulnerabilities and privacy issues. This paper explains four cyberattacks and a case scenario of each attack as it relates to the metaverse. Additionally, this study developed a metaverse forensic framework that can be used to investigate cyberattacks in the metaverse world. Furthermore, this study describes how forensic examiners can conduct a forensic investigation using state-of-the-art forensic solutions and tools. The developed framework can be used by forensic examiners, security researchers, as well as the general scientific community for the security of the metaverse.

Keywords: metaverse forensics, metaverse security, virtual world forensics, metaverse forensic framework.

1. Introduction

Digital forensics is the aspect of information assurance that entails the proactive and reactive process of understanding the who, how, what, and how an event occurred in a digital medium with a probable application in litigation (Adeyemi et al., 2013; Al-Dhaqm et al., 2021; Ellison et al., 2019). Given that digital mediums play a critical role in human daily living, it suffices to highlight that the increasing crime rate provides a frequent need for digital forensics. Recent technological advances illustrate the insatiability of users, with a more targeted drift from physical to virtual reality. Virtual reality (VR) is a technology that enables immersion in a digital environment. It provides promising opportunities for teaching and learning, particularly for the transfer of practical skills. This technology offers benefits, such as time and cost savings, as well as increases in training effectiveness and safety in industries including manufacturing, construction, and healthcare (Ghobadi, Mohsen and M.E. Sepasgozar, 2020). To aid flexibility, a VR world can be placed on removable media which implies that the installed operating systems reside on an external storage device. System administrators have found these tools to be very helpful, as it aids the projection of physical quantity into an intangible environment while maintaining all elements of a physical environment.

The notion of the virtuality continuum has also been extended to include augmented reality (a process in which digital information is overlaid in the real world) and mixed reality. An illustrative depiction of the relationship between these terminologies is further highlighted in Figure 1. Mixed reality (MR) can be expressed as the combination of augmented and virtual reality. In the context of the Metaverse, mixed reality refers to the integration of MR technologies into the metaverse, allowing users to experience and interact with the virtual world in a more immersive and interactive way. In recent years, the application of VR technology has increased across several sectors (Ghobadi, Mohsen and M.E. Sepasgozar, 2020). One of these is the construction sector, which has the potential to provide stakeholders with an immersive experience of the intended architectural masterpiece. It also provides the potential to identify risks and faults in design to some stakeholders (Ghobadi, Mohsen and M.E. Sepasgozar, 2020). Similarly, the entertainment and gaming industry share great benefits from VR as it provides a better interactive platform for user immersive experience. However, with these advantages comes a high probability of misuse. The VR platform can be leveraged to conduct several cyber-related attacks including but not limited to insider misuse, espionage, information leakage, impersonation attack, cyber sabotage, misinformation dissemination, cyberstalking, information theft, as well as covert recruitment.

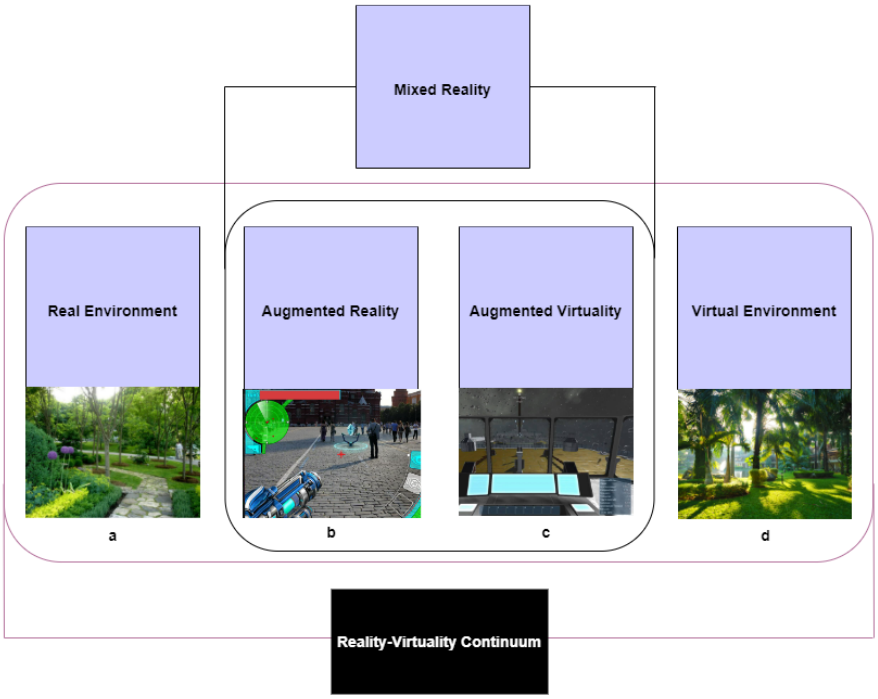


Figure 1: Reality-Virtuality Continuum that Indicates the Stages Between Real and Virtual Environment (Ghobadi, Mohsen and M.E. Sepasgozar, 2020)

However, the list of potential crimes will only grow as more people use VR technology and its associated advances (Aloqaily et al., 2022; Chow et al., 2023; Qin et al., 2022; Wang et al., 2022). This may further include financial fraud, ransomware, and phishing (INTERPOL, 2022; Odeleye et al., 2023; Vondrek et al., 2022). Whilst the Metaverse provides some inherent mechanisms for preventing some attacks, through deterrence and detection, some of these risks could be relatively difficult to investigate. A probable justification for this assertion is that both legal and illegal actions could be carried out using the same content, albeit, with a different context. Also, a virtual world requires some degree of creativity which is unique to each platform. Thus, some actions which are considered illegal in the physical environment may be accepted in some Metaverse platforms, as an immersive experience. Therefore, not all malicious act in the real world is illegal in the virtual environment. VR technology is one of the technologies used in the Metaverse experience. Any other technology that replicates or improves real-world experiences with technology may be referred to as the metaverse. According to GlobalData ("The Outlook of Metaverse Market 2022-2030," 2022), from 2022 to 2030 the worldwide metaverse market would increase from USD 22.79 billion to USD 996.42 billion, with a compound annual growth rate (CAGR) of 39.8%. It is the upcoming major trend in digital media, while VR and AR are key technologies fostering its growth ("The Outlook of Metaverse Market 2022-2030," 2022).

Metaverse is being used nowadays dramatically in advanced blockchains where Bitcoin, Ether, and Dogecoin are all powered by blockchain technology. The advanced blockchain facilitates the development of digital assets known as non-fungible tokens (NFTs) and applications while serving as a distributed ledger for recording peer-to-peer transactions (LeewayHertz, 2022). Other than that, Businesses are moving away from the two-dimensional surface of e-commerce and embracing lifelike virtualized worlds for a meaningful experience thanks to the growing application of the Metaverse. Owners of e-commerce businesses may conduct trade formalities including product inspection, negotiations, and transaction closure with merchants in a virtual setting. Additionally, rather than depending on digital marketing strategies, companies may better impact customers by creating engaging and realistic marketing material (LeewayHertz, 2022). However, recent statistics revealed that the metaverse poses diverse dangers to Internet users. This includes addiction, mental and physical health challenges, and privacy issues. A breakdown of the distribution is further presented in Figure 2.

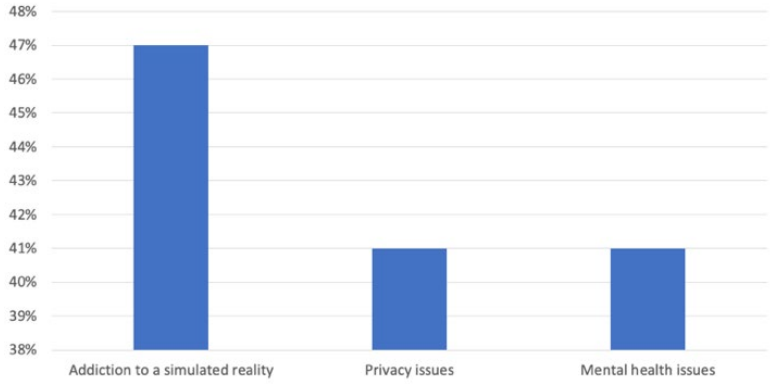


Figure 2: Dangers of the Metaverse According to Internet Worldwide in 2021 (Statista, 2022)

As illustrated in Figure 2, Users and society at large may experience certain unforeseen effects of the metaverse. 47% of the sampled population asserts that addiction to a virtual or simulated environment is a potential menace that should be adequately remediated. Furthermore, given that there is a need to investigate crimes associated with digital platforms, law enforcement, and digital forensics units must create and maintain an efficient quality assurance system to ensure that potential digital evidence (PDE) is gathered, kept, reviewed, or transmitted in a way that protects the correctness and dependability of the PDE. Standard operating procedures (SOPs) and/or models that adhere to the chain of custody are the initial component of this system. These rely on the digital forensics "process-phase-procedure-task-subtask" sequence (Bulbul et al., 2013). To the best of the Authors' knowledge, this is the first study that attempts to provide a typical SOP for conducting a digital investigation in a virtual platform using potential cyber-attacks. The remainder of this manuscript is structured as follows: in the next section, the methodology adopted in this study is presented. This is followed by a hypothetical case study and the result of the experimental process. A concise discussion and potential future works are thereafter provided.

2. Methodology

This study leverages an experimental approach to metaverse framework development. To do so, the operational framework shown in Figure 3 is followed. The operational framework comprises 3 phases which are further explained.

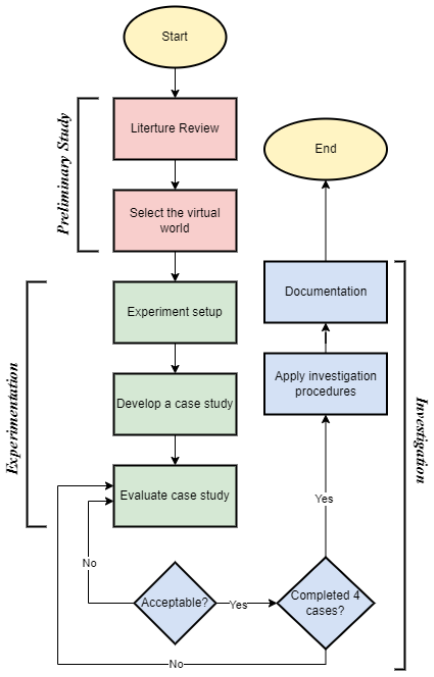


Figure 3: Operation Framework Flow Model

Given that the goal of this paper is to develop a metaverse forensic framework that examiners can follow to be able to investigate a metaverse case. This was achieved by conducting an extensive literature review where a particular category of metaverse was selected, which is Virtual Reality (VR), precisely the fully immersive virtual reality. This virtual world of interest was selected because it offers a state-of-the-art on metaverse. Virtual reality, or VR, is a technology that creates a simulation of a 3D environment that enables the user to interact and explore a virtual surrounding that simulates a real-world experience through the senses of the user ("The Outlook of Metaverse Market 2022-2030," 2022). After the technology of the VR was selected, an experimental setup was established. The experimental setup includes setting up an account in Oculus VR and integrating (wired or wirelessly) a Meta Quest VR headset. Steps to create an Oculus account and interfacing the Meta Quest VR headset to a workstation are further provided in Figures 3 and 4. The workstation comprises a core i7 10th generation, 32GB RAM, and runs on a Microsoft Windows 10 Professional. To ascertain the experiment process, case studies were developed and evaluated. Each case study describes a cyberattack scenario applicable within the virtual reality world. When a case study was created, it was reviewed iteratively until a forensic investigation process can be initiated. This implies that each attack aligns with Locard's exchange principle. Using this heuristic selection approach, a total of four cases were developed, upon which the investigation and documentation stages are undertaken. The result of the respective case studies is presented in the proceeding section.

3. Result and Analysis

To enter the metaverse, users will take the form of digital avatars that are complemented with VR/ AR technologies that help in ensuring a sense of existence in the metaverse. The user can control the environment, look around at any scene, or interact with the objects in the scene using the headset. Newer headsets such as the Oculus Rift, PlayStation VR, and Samsung Gear VR tend to provide better interactivity. The Meta Quest VR headset is adopted in the first scenario. The process model provided in Figure 4 describes the process of using this headset. This is a stepwise process required to access the Oculus account.

After creating an account, the Meta Quest application was installed on the workstation, and user access was granted by logging into a newly created account. Once logged in, the VR Meta Quest headset can be connected in two ways: wired or wirelessly. The process model presented in Figure 5 was used to establish the VR headset connectivity in this study, both in wired and wireless modes. These steps were adapted for this study as they can enhance experimental repeatability. To substantiate the need for a forensic investigation framework, this study leverages an attack approach to scenario development. These are presented in the next subsection.

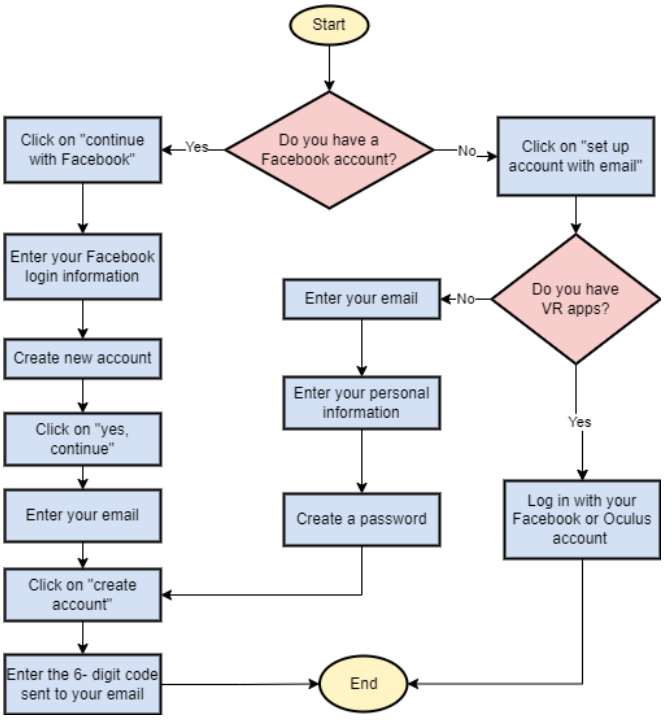


Figure 4: Signing Up for Oculus Account Flow Model

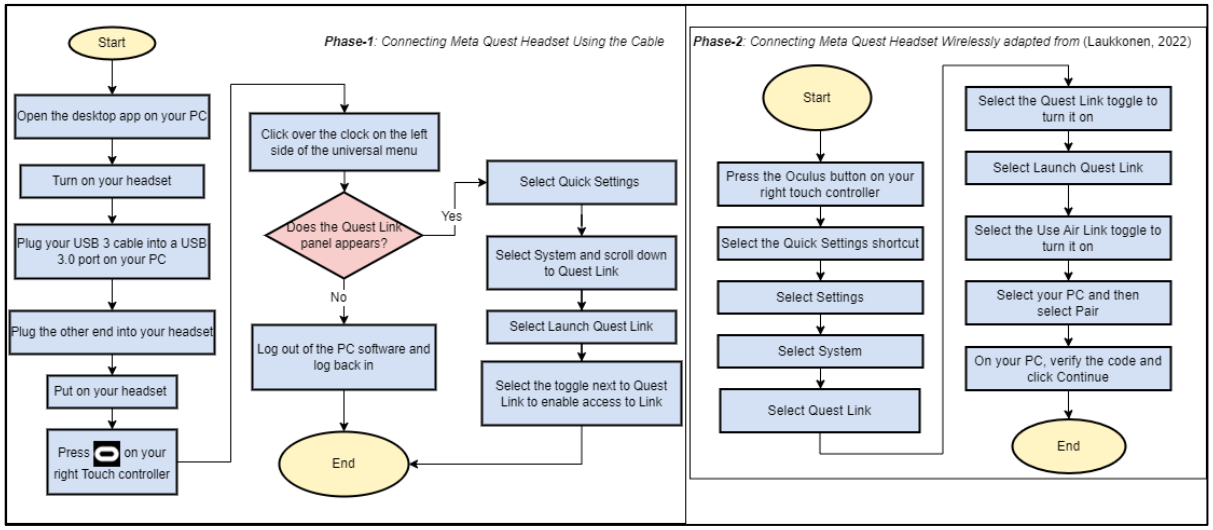


Figure 5: MetaQuest Headset Process model

3.1 Downloading Malicious Codes

An employee, Mr. Mark, for Globe-Star PLC is responsible for creating an interactive VR environment for entertainment purposes. Thus, he used a special workstation PC in the company to connect his Head-Mounted Device (HMD) to download his special software. However, one day the IT team informed him that his PC was involved in downloading malicious codes that are embedded VR based games. Mr. Mark told them that he didn't download this application and he was only doing his job. To investigate such a case, forensic investigators can prove that Mark was not the one who downloaded this application by first collecting the evidence, the VR headset and the computer. Second, secure the evidence by storing it in a secure forensic lab. Third, collect information about the main user and the tools that were used as well as the software. Fourth, collect data by connecting the VR headset to a computer and running String VR on it. Using these steps, the software will generate four reports:

1. *Device.txt*: stores the list of devices that are connected and other important information, for example, which connected ports.
2. *Logs.txt*: contains the logs of different users with the accurate timing, as well as any inside and outside processes that transverse outside and inside the network.
3. *Main.txt*: contains general information about the software used and how devices connect to it, what time, which port, and other information.
4. *Raw.txt*: stores raw information that is not yet processed to give specific information.

Fifth, extract data using FTK Imager. This program will allow them to facilitate data viewing and imaging. The output of the analysis is further shown in Figures 6 and 7.

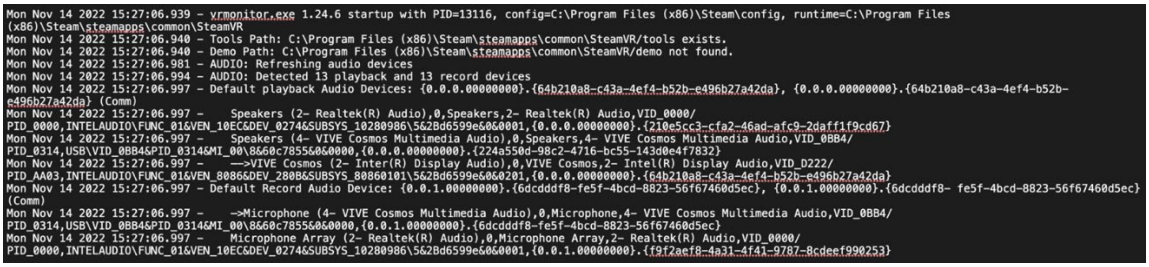


Figure 6: Devices Used to Access the Software

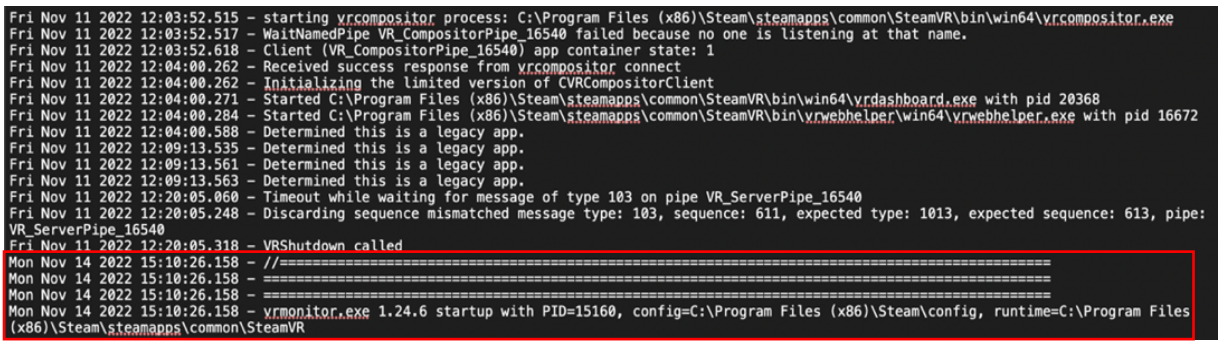


Figure 7: The Device Logins by Mr. Mark

The log file displays the specific date and time Mark used the software and how he used it. The evidence shows that he was inside the workstation place on the dates: 11/11/2022 and 14/11/2022. According to the permission granted to Mark, he usually accessed the workstation on Monday and Friday only. On 15/11/2022, an unknown user was found, and the user created an account with the profile name *8000****. The result of the FTK image allowed for the discovery of all downloaded games during the days that Mark wasn't in the workstation (Date →15/11/2022). The device that was used is presented in Figures 8a and 8b. Furthermore, the log file, Figure 8c, showed that there was a log attempt between 5:00 pm and 6:00 pm on 15/11/2022. This corroborative evidence can be used to support the claim of Mr. Mark, as he was not within the premise on the stipulated day.

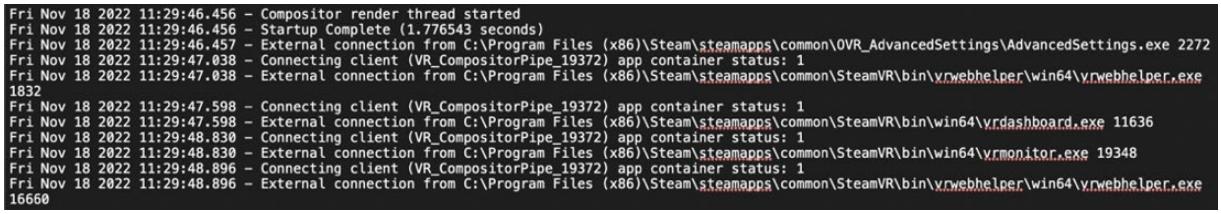


Figure 8a: Date and Time Mr. Mark Connected and Accessed the Workstation.

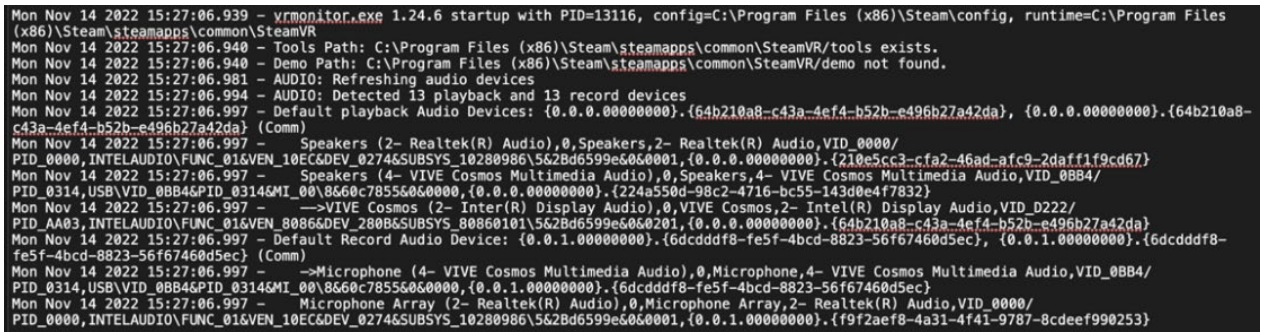


Figure 8b: Devices Used

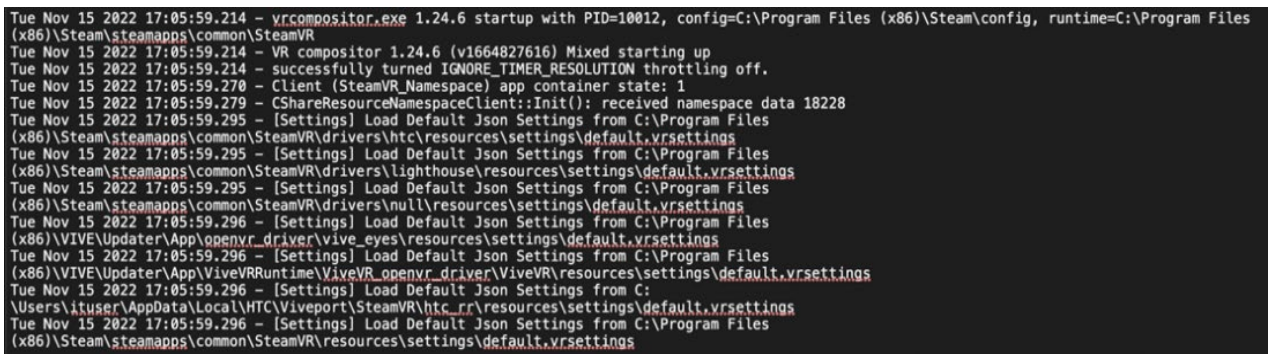


Figure 8c: Date and Time of Mr. Mark's Account Logins.

As a result, the forensic examiners can conclude that Mr. Mark is not involved in the misuse raised against him because the login and the application downloaded prove that Mr. Mark was not available in the vicinity during the time of the incident. In addition, the person who accessed Mr. Marks' computer left traces of information such as a username that could be used for attribution.

3.2 Case 2: Impersonation

JohnSmith and his friend TashaCoin share all online login credentials irrespective of the platform. Both of them have the Oculus VR headset and their account credential. As JohnSmith was attempting to login into his oculus account, he saw TashaCoin looking at the keyboard but he didn't say anything because he trusts his friend. One day, JohnSmith accessed his account and found some strange applications on his PlayStation account. Furthermore, some strange messages with unknown entities that relate to bullying were also found in the account. Given that JohnSmith did not carry out these actions, and TashaCoin claimed ignorance, there was a need to investigate this case to gain insight into the perpetrator. To do so, a forensic examiner begins by logging into JohnSmith's Oculus account using this link https://auth.meta.com/. Once logged in, navigate to settings>Account > view your information > Security and login information. In this directory, the dataset for event correlation can be extracted from the account. The information typically includes active sessions, location history, devices used, recently viewed items, login history, as well network-related potential digital evidence (PDE). As a PDE, the network parameter of interest includes IP address, MAC address, and well geolocation data. The experimental approach and outcome are further elucidated:

1. We logged into the Oculus account through this link: https://auth.meta.com/
2. Navigate to Settings > Account > View your information > Security and login information. The output of the experimental process is further presented in Figure 9.

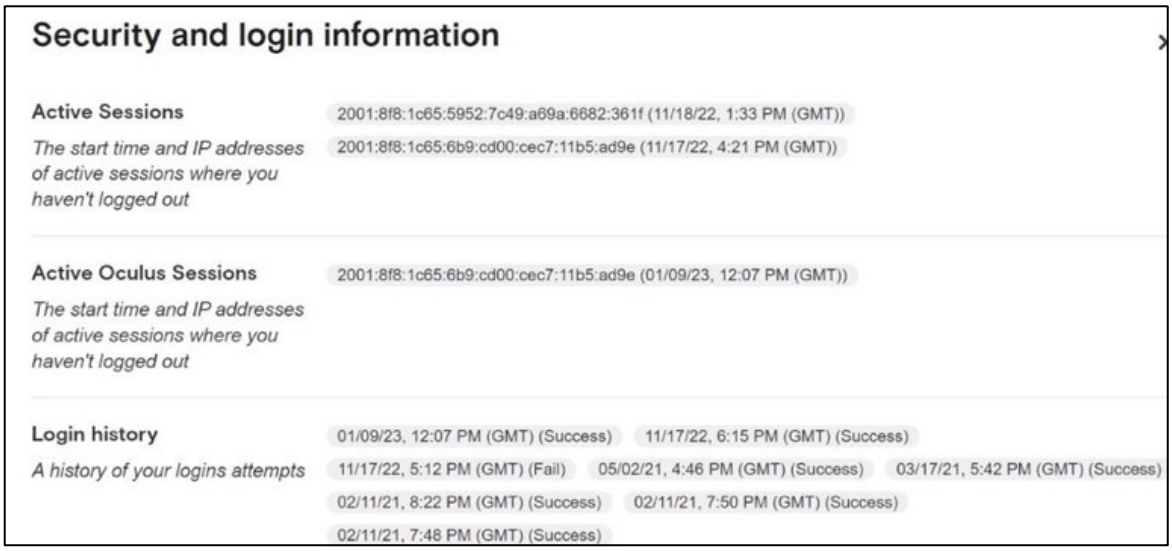


Figure 9: Result of the Experimental Process of Impersonation Investigation

3. Two IPV6 addresses are displayed in the active session which indicates that two people logged in to this account.
4. Investigating further, the command "ipconfig/all" was instantiated on the workstation terminal. It revealed the IPV6 address of 2001:8f8:1c65:6b9:cd00:cec7:11b5:ad9e.
5. To confirm this, the Ipv6 of all suspects was analyzed. This revealed IP address of TashaCoin was 2001:8f8:1c65:5952:7c49:a69a:6682:361f.
6. This showed that TashaCoin logged into JohnSmith's account on 11/18/2022 at 1:33 PM (GMT).

3.3 Case 3: Data Leakage

There is a student named Asma who owns a VR headset. One day she decided to bring her headset to the university so her friends can play with it. After they were done playing, they left the PC running and went to the cafeteria to get some food. However, they spent hours eating and the PC was still unlocked. When they got back their friend told them that Asma's information was leaked including her credit card details. Asma realized that

someone used the PC and gone through her account while they were gone. She looked around and saw a lot of students in the lab and when she asked them who used the PC, they all said that three students used that computer. She was unable to know which one did it as all of them claimed that they were just finishing their university work. Forensic examiners can investigate the source of the leak by capturing and examining the keystroke forensics and the mouse dynamics of the computer. This will allow them to collect data about the actions that were done on the computer. Moreover, they will be able to know which pattern belongs to which suspect. Once the data is collected and they know when the attack happened, they can go to the keystroke database and extract the keystroke and mouse dynamics and analyze them. Moreover, they can now view the surveillance camera video and create the timeline of the attack. Once they reach the time of the attack, they will be able to know which student leaked Asma's information as they will appear on the video.

3.4 Case 4: Malware

Sheikha has an Oculus Rift VR headset that she uses all the time. One day when she tried to start up her PC, all her files were corrupted. She realized that an attacker might have installed malware on her computer as she clicked on a suspicious link yesterday while she was using her Oculus account. Sheikha started to get worried as all her files on the computer were important. Forensic examiners can conduct their investigation by going through the event log, and network log, and checking the running processes. As a result, they will be able to collect evidence and see how the attacker was able to install the malware and track him down. A summary of the overall forensic usefulness of the explored case studies is provided in Table 1.

Table 1: Forensic Usefulness of the Metaverse Case studies

Case studies	Attack Type	Potential digital forensic evidence location	Nature of the potential evidence
Case - 1	Overlay attack	Oculus VR headset	non-volatile
Case - 2	Impersonation	Oculus account settings	non-volatile
Case - 3	Data leakage	Computer database (keystroke and mouse dynamics) and the surveillance camera	non-volatile
Case - 4	Malware	Event log, and network log, and checking the running processes	non-volatile

4. Discussion and Recommendation

With reasonable certainty, it can be asserted that all VR systems are vulnerable to these types of attacks as explored in the respective cases. With such technological improvement, there is a greater need for enhanced security. As asserted in existing studies (Odeleye et al., 2023; Vondrek et al., 2022), the metaverse presents a platform for enhanced user interactivity which can be leveraged for several nefarious activities. With the growing adoption of the metaverse, especially in the e-commerce industry, it is essential to develop a threat-based use-case scenario for metaverse security. Given that the metaverse could transcend traditional online interaction into a more personal immersive ecosystem, the potential consequence of malicious attacks could lead to psychological and physical defects. Whilst the current study targeted the Oculus VR system, the Authors believe that the attacks identified in the cases can happen to all VR systems. The security of the VR and its safety features need further protection and improvement. Although there are some benefits in having the user manually enter their private information to create their account, gaining access to this data should be guarded. The user's data should be protected so that the attacker cannot have access to the user's data.

To protect the data, this study found that it is better to encrypt such data as it was revealed that some data were stored in plaintext. Moreover, access to the services should be restricted, as this will aid integrity preservation and data usefulness. Virtual world data is vulnerable to many attacks in comparison with standard screens given that virtual reality (VR) offers new ways to perceive data. Also, it is inherently portable and offers more connectivity possibilities. Existing findings and the implications found in security studies include manipulatable 3D objects used for authentication, which can be correlated to the considerable cognitive effort needed to simultaneously monitor many visual channels, including hand motions while seeing the manipulations (Mathis et al., 2020). The metaverse world can exploit several opportunities in the coming years, as it requires

multisensory interactions with several technologies: virtual environments, digital objects, and people immersive technologies, virtual reality (VR), augmented reality (AR), mixed reality (MR), and extended reality (XR) (Trunfio & Rossi, 2022). These technologies support the creation of the metaverse and facilitate immersive experiences in the digital world. Virtual reality technology provides users with a connected experience in the metaverse. On the other hand, augmented reality expands the use of virtual reality by overlaying digital information onto the physical environment. XR is an extended reality, a term used to include VR, AR, and MR. XR is used for virtual commerce or v-commerce to create computer-mediated indirect experiences (Trunfio & Rossi, 2022). With the growing need for virtual reality, specifically with the Metaverse, it is evident that cybercrime will evolve.

The findings from this current reveal the relevance of logs and systems documentation. This includes the device log, as well as the connectivity log. As highlighted in Table 1, these logs could be leveraged for a different context. Furthermore, from a forensic standpoint, it will be helpful to forensics examiners as they will know how they could conduct their investigation based on what attack happened. This also relates to the reliability and admissibility of forensic evidence. Understanding that some content of the VR headset can be manipulated by an adversary presents a logic for further analysis of results obtained from the virtual world. Forensic examiners would then be required to provide a suitable means to identify the users and corroborate claims where applicable. Studies have alluded to the need for a forensic readiness approach to volatile environments (Lagrasse et al., 2020; Munkhondya et al., 2019, 2020) to ensure potential evidence availability and reliability. This is specifically applicable in the virtual reality platform where potential evidence might not be available after the occurrence of the incident. Whilst the logs considered in this current study relate to non-volatile PDE, this might not be the case for some special cases. For instance, a situation whereby an ongoing attack is to be investigated while the malicious actor is still using the platform. Therefore, the findings from this study provide a veritable process model for investigating incidents in the metaverse, based on the devices used. The finding in the study is, however, limited to log-related evidence correlation and event corroboration. Further study will be carried out on the feasibility of developing a real-experience metaverse crime which could leverage approaches such as impersonation of known users, identity theft, as well as false information dissemination. These attacks are postulated to require a near-real-time potential digital evidence collection and event correlation. This is, in addition, to the observable logs. Furthermore, the exploration of such processes would lead to the development of a forensic readiness mechanism for metaverse forensics. The latter observation can complement the find in this study, but developing a process for a forensically ready log-generation mechanism.

5. Conclusion and Future Works

Metaverse technology employs avatars and infrastructure, platforms, and gadgets that allow immersive experiences, and create a mixed-reality environment where people and objects may interact synchronously and live outside of time and space. Virtual reality is a useful technology that helps people interact with each other. Though platforms such as Metaverse create an environment for interaction for people who are worlds apart, there are cybersecurity-related risks associated with such platforms which may compromise the privacy of users on such platforms. This study explored how forensic investigators can leverage various logs within the metaverse platform to conduct an investigation. However, in the case of privacy violations and other forms of cyberattacks, Digital forensics tools such as FTK Imager can be used to investigate and trace the specific user actions that led to the attack/violation, which could uncover the real culprit for legal actions.

References

Adeyemi, I. R., Razak, S. A., & Azhan, N. A. N. (2013). A review of current research in network forensic analysis. *International Journal of Digital Crime and Forensics*, *5*(1), 1–26. https://doi.org/10.4018/jdcf.2013010101

Al-Dhaqm, A., Ikuesan, R. A., Kebande, V. R., Razak, S. A., Grispos, G., Choo, K. K. R., Al-Rimy, B. A. S., & Alsewari, A. A. (2021). Digital Forensics Subdomains: The State of the Art and Future Directions. *IEEE Access*, *9*, 152476–152502. https://doi.org/10.1109/ACCESS.2021.3124262

Aloqaily, M., Bouachir, O., Karray, F., Ridhawi, I. Al, & Saddik, A. El. (2022). Integrating Digital Twin and Advanced Intelligent Technologies to Realize the Metaverse. *IEEE Consumer Electronics Magazine*, 1–7. https://doi.org/10.1109/MCE.2022.3212570

Bulbul, H. I., Yavuzcan, H. G., & Ozel, M. (2013). Digital forensics: An Analytical Crime Scene Procedure Model (ACSPM). *Forensic Science International*, *233*(1–3), 244–256. https://doi.org/10.1016/J.FORSCIINT.2013.09.007

Chow, Y., Susilo, W., Li, Y., Li, N., & Nguyen, C. (2023). *Visualization and Cybersecurity in the Metaverse : A Survey*.

Ellison, D., Ikuesan, R. A., & Venter, H. S. (2019). Ontology for Reactive Techniques in Digital Forensics. *2019 IEEE Conference on Application, Information and Network Security, AINS 2019*, 83–88. https://doi.org/10.1109/AINS47559.2019.8968696

Ghobadi, Mohsen and M.E. Sepasgozar, S. (2020). Smart Cities and Construction Technologies. In *Smart Cities and Construction Technologies*. https://doi.org/10.5772/intechopen.86103

INTERPOL. (2022). *INTERPOL launches first global police Metaverse*. International Criminal Police Organization. https://www.interpol.int/en/News-and-Events/News/2022/INTERPOL-launches-first-global-police-Metaverse

Lagrasse, M., Singh, A., Munkhondya, H., Ikuesan, A., & Venter, H. (2020). Digital forensic readiness framework for software-defined networks using a trigger-based collection mechanism. *Proceedings of the 15th International Conference on Cyber Warfare and Security, ICCWS 2020*, 296–305. https://doi.org/10.34190/ICCWS.20.045

Laukkonen, J. (2022). *How to Connect Meta (Oculus) Quest 2 to a PC Wirelessly*.

LeewayHertz. (2022). Metaverse Use Cases and Benefits. In *LeewayHertz*.

Mathis, F., Williamson, J., Vaniea, K., & Khamis, M. (2020). RubikAuth: Fast and Secure Authentication in Virtual Reality. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3334480.3382827

Munkhondya, H., Ikuesan, A. R., & Venter, H. S. (2020). A case for a dynamic approach to digital forensic readiness in an sdn platform. *International Conference on Cyber Warfare and Security*, 584--XVIII.

Munkhondya, H., Ikuesan, A., & Venter, H. (2019). Digital Forensic Readiness Approach for Potential Evidence Preservation in Software-Defined Networks. *14th International Conference on Cyber Warfare and Security*, 268–277. https://search.proquest.com/docview/2198531158?accountid=169469

Odeleye, B., Loukas, G., Heartfield, R., Sakellari, G., Panaousis, E., & Spyridonis, F. (2023). Virtually secure: A taxonomic assessment of cybersecurity challenges in virtual reality environments. *Computers and Security*, *124*. https://doi.org/10.1016/j.cose.2022.102951

Qin, H. X., Wang, Y., & Hui, P. (2022). *Identity, Crimes, and Law Enforcement in the Metaverse*. http://arxiv.org/abs/2210.06134

Statista. (2022). Benefits of the metaverse worldwide 2021. *Statista Research Department*. https://www.statista.com/statistics/1285117/metaverse-benefits/

The Outlook of Metaverse Market 2022-2030. (2022). In *GlobalData*.

Trunfio, M., & Rossi, S. (2022). Advances in Metaverse Investigation: Streams of Research and Future Agenda. *Virtual Worlds*, *1*(2), 103–129. https://doi.org/10.3390/virtualworlds1020007

Vondrek, M., Baggili, I., Casey, P., & Mekni, M. (2022). Rise of the Metaverse's Immersive Virtual Reality Malware and the Man-in-the-Room Attack & Defenses. *Computers & Security*, *September*, 102923. https://doi.org/10.1016/j.cose.2022.102923

Wang, Y., Su, Z., Zhang, N., Xing, R., Liu, D., Luan, T. H., & Shen, X. (2022). A Survey on Metaverse: Fundamentals, Security, and Privacy. *IEEE Communications Surveys and Tutorials*, 1–32. https://doi.org/10.1109/COMST.2022.3202047

Attention-Based Deep Learning Modelling for Intrusion Detection

Ban AlOmar[1], Zouheir Trabelsi[2] and Firas Saidi[3]
[1,2]College of Information Technology, United Arab Emirates University, Al-Ain, UAE
[3]Department of Computer Science, SMPCS, University of Reading, UK

700039223@uaeu.ac.ae
trabelsi@uaeu.ac.ae
f.saidi@reading.ac.uk

Abstract: Cyber-attacks are becoming increasingly sophisticated, posing more significant challenges to traditional intrusion detection methods. The inability to prevent intrusions could compromise the credibility of security services, thereby putting data confidentiality, integrity, and availability at risk. In response to this problem, research has been conducted to apply deep learning (DL) models to intrusion detection, leveraging the new era of AI and the proven efficiency of DL in many fields. This study proposes a new intrusion detection system (IDS) based on DL, utilizing attention-based long short-term memory (AT-LSTM) and attention-based bidirectional LSTM (AT-BiLSTM) models. The time-series nature of network traffic data, which changes continuously over time, makes LSTM and BiLSTM particularly effective in handling intrusion detection. These models can capture long-term dependencies in the sequence of events, learn the patterns of normal network behaviour, and detect deviations from this behaviour that may indicate an intrusion. Also, the attention mechanism in the proposed models lets them make predictions based on the most important parts of the network traffic data. This is important for finding intrusions because network traffic data can have many different features, not all of which are important for finding an attack. The attention mechanism lets the models learn which features are most important for making accurate predictions, which improves their performance and efficiency. The UNSW-NB15 benchmark dataset is used in the study to measure and compare the effectiveness and reliability of the proposed system. This dataset contains normal and attack traffic data with a significant class imbalance. To address this issue, the study employs the Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset, thus reducing the risk of overfitting to the majority class and improving the model's performance in detecting attacks. The performance evaluation results demonstrate that the proposed models achieved a detection rate of over 93%, indicating high precision in detecting intrusions. By harnessing the power of deep learning, these models can learn and adapt to new threats over time, thus ensuring data confidentiality, integrity, and availability in today's interconnected world.

Keywords: Intrusion Detection, Attention Architecture, BiLSTM, LSTM, Network Attacks

1. Introduction

With the increasing frequency of threats and attacks on computer networks, ensuring the effectiveness of intrusion detection systems (IDS) has become difficult. The present-day systems are heterogeneous, dynamic, and complex, making it challenging to address multiple attacks. Based on the behaviour of intrusion detection, network intrusion detection can be divided into two categories: anomaly-based IDS and signature-based IDS (SIDS) (Zhang *et al.*, 2022). The anomaly-based approach detects new attacks by creating a model of normal activities within a system and identifying potential attacks from behaviours that deviate from the established normal behaviour pattern. In contrast, SIDS uses matching techniques to detect known attacks. Whenever a behaviour signature matches one from a previous intrusion existing in the signature database, the behaviour is classified as an intrusion (Bakhsh *et al.*, 2019).

The IDS's effectiveness is evaluated based on its ability to detect actual intrusions accurately. In practice, the IDS's efficacy is determined by its ability to minimize the false-positive rate (FPR) instead of maximizing the true-positive rate (TPR), as described in (Bakhsh *et al.*, 2019).

However, current IDS face several significant limitations. For example, anomaly detection IDS sometimes classifies rare normal behaviours as abnormal. At the same time, the diversity in the implementation mechanisms of the operating systems makes building a unified pattern library very difficult, which decreases the efficiency of the SIDS (Sarhan *et al.*, 2021).To overcome these limitations, various ML and DL techniques have been studied to design IDS. These techniques act as classifiers to determine whether events are attacks or benign occurrences and to specify the kind of attacks. When it comes to managing massive volumes of data and developing models that can be applied to different network environments and attack vectors, DL outperforms standard IDS (Mighan and Kahani, 2021).

The goal of this paper is to show how AT-LSTM and AT-BiLSTM can be used in intrusion detection. The key contributions of this paper are:

- An effective and compact deep learning model is presented that makes use of attention-based LSTM and attention-based BiLSTM architectures. Network traffic data used for intrusion detection is typically sequential, with each network packet being captured as a series of features across time. LSTM and BiLSTM networks, which are intended to handle variable-length input sequences, may efficiently simulate this sequential data.
- Synthetic Minority Oversampling Technique (SMOTE) is used as an oversampling strategy to address the problem of unequal class representation in the dataset.
- Evaluation of the proposed models' performance using the F1-score, accuracy, precision, recall, and the common UNSW2015-NB15 network security dataset.

The structure of this paper is as follows: Section 2 provides an overview of prior research related to intrusion detection using ML and DL techniques. Section 3 outlines the model's design. Model evaluation and discussion are presented in Section 4. The paper concludes with the key findings of our work in Section 5.

2. Related Works

There has been a lot of research on how ML and DL might be employed to increase intrusion detection. This section highlights the research that has applied ML and DL to intrusion detection. A stacked autoencoder network was used by the authors in (Mighan and Kahani, 2021) to extract features before the data was classified using support vector machines, random forests, decision trees, and Naive Bayes. The UNB ISCX 2012 dataset was used to validate the model, and the maximum degree of accuracy was 90.2%. The authors of (Liu, Gu and Wang, 2021) developed a binary classification model using K-means and Random Forest (RF). They used Convolutional Neural Networks (CNN) and LSTM to categorise attacks into different groups. The effectiveness of their model was evaluated on two benchmark datasets, with accuracy rates of 85.24% and 99.91%. GRU and improved LSTM were used by A. Aldallal to build an IDS (Aldallal, 2022). Using information from the CICIDS 2018 dataset, the model was evaluated. The model outperformed the benchmarks by 12.045%. The author (Xu *et al.*, 2018) used GRU and Multi-Layer Perceptron (MLP) to build an IDS model and tested its effectiveness on the KDD-99 and NSL-KDD datasets. The model achieved a prediction rate of 99.42% and 99.31% on KDD-99 and NSL-KDD, respectively. Using LSTM, the author of (Yan *et al.*, 2020) built an IDS using a stacked autoencoder. The model is tested on KDD Cup-99 dataset, demonstrating that it outperforms several conventional ML models. The author (Benaddi *et al.*, 2022) suggested a Deep Reinforcement Learning -based IDS . The NSL-KDD dataset was used to evaluate the proposed DRL-performance IDS. In (Gaifulina and Kotenko, 2021), the authors evaluated deep neural network models for anomaly detection in Internet of Things (IoT) network traffic. They evaluated the quality metrics of anomaly detection as well as the amount of time spent training the models using the UNSW-NB 15 dataset. The author of (Yang *et al.*, 2022) used an intrusion detection model with a multilayer attention mechanism for a power information network. Compared to the no-attention layer, the attention layer increased the detection rate by 1.99%. In (Wei *et al.*, 2023), the efficiency of an attention mechanism and an autoencoder for intrusion detection was investigated. The suggested model outperformed conventional machine learning techniques in tests utilising a dataset of actual CAN bus messages from in-vehicles. An attention-based LSTM neural network was also used in a study reported by (Liu *et al.*, 2020) to forecast air pollution. In (Lan *et al.*, 2020), a threshold-optimized CNN-BiLSTM-Attention technique was suggested to address the issues of class imbalance and low detection rate in minority classes. The suggested method combines the CNN-BiLSTM-Attention model with a threshold adjustment technique based on the ROC curve. The technique was tested using an industrial data set proposed by Mississippi State University in 2014, and the findings showed an accuracy of 96.7%. The author of (Laghrissi *et al.*, 2021) suggested a detection model that combines four feature reduction algorithms—Chi-Square, UMAP, principal components analysis (PCA), and mutual information with LSTM and attention mechanisms. The method was tested on the NSL-KDD dataset and acheived an accuracy of 99.09% and 98.49% for binary and multiclass classification, respectively. In (Cao *et al.*, 2021), the authors suggested a Dense Attention-LSTM (DAL) IDS. The suggested model uses a CuDNN-based LSTM network to learn time-related information about traffic, global max-pooling to compress data and enhance generalisation capabilities, attention mechanism to capture key features, dense dilated convolutions to extract underlying features of network traffic. The suggested model achieved a binary classification accuracy of 92.65% and recognised different attacks with an accuracy of 81.28%.

3. Methodology

This section describes the steps needed to build the proposed intrusion detection system. These steps include pre-processing the data, choosing a model, training the model, evaluating the model, and analysing the results.

3.1 Dataset

The UNSW-NB 15 dataset (Moustafa and Slay, 2016; Sarhan *et al.*, 2021) includes a combination of realistic normal activities and synthetic attack behaviours generated from network traffic. UNSW-NB15 is one of the most enormous and widely used benchmark datasets for evaluating intrusion detection systems (Sarhan *et al.*, 2021). The dataset has over two million network connection records and a wide range of attack types, such as backdoors, denial of service (DoS), exploits, fuzzers, generic, normal, reconnaissance, shellcode, and worms.

3.2 Data Pre-processing

For ML to work well, the data must be properly prepared. This means taking out numerical and statistical features, reducing sparsity, and making the data fit for machine learning algorithms. In the case of the UNSW-NB15 dataset, missing values were taken out, categorical features were turned into numerical values using one-hot encoding, and features were normalised using Min-Max scaling. The UNSW-NB15 dataset had a problem with a class imbalance. Some attack types were much less common than the "normal" class. Specifically, the "worms," "backdoor," "shellcode," and "analysis" attack types were the minority classes with the smallest number of samples. SMOTE was used to handle the class imbalance in the dataset. This ensures that the model gets a more balanced exposure to all classes during training and can improve the overall performance of the model (Zhang *et al.*, 2018; Waqar *et al.*, 2021).

3.3 Attention Mechanism

The attention layer is part of a neural network architecture that can be used to pay more attention to certain parts of the input sequence. The basic idea is to compute a weight for each element of the input sequence based on how relevant it is to the problem. The equations for the attention layer can be written as follows (Zhang *et al.*, 2019; Liu *et al.*, 2022).

Let $h_1, h_2, \ldots, h_T$ be the input sequence of hidden states, where $h_t \in R_d$ for t=1, 2, …, T and d is the dimensionality of the hidden state. First, the attention scores α_t between the current hidden state of the network and each element of the input sequence are calculated using the following equation:

$$\boldsymbol{\alpha_t} = \frac{\boldsymbol{exp(e_t)}}{\sum_{j=1}^{T} \boldsymbol{exp(e_j)}} \quad \textbf{(1)}$$

Where, e_t is the relevance of the t^{th} element of the input sequence, defined by some function of the current hidden state of the network and the t^{th} hidden state of the input sequence:

$$\boldsymbol{e_t} = \boldsymbol{f}\left(\boldsymbol{h_t}, \widetilde{\boldsymbol{h}}\right) \quad \textbf{(2)}$$

Where $\hat{h}$ is a summary vector of the hidden states, computed using a weighted sum:

$$\widetilde{\boldsymbol{h}} = \sum_{i=1}^{T} \boldsymbol{\beta_i h_i} \quad \textbf{(3)}$$

And β_i is a weight assigned to each hidden state h_i, computed as follows:

$$\boldsymbol{\beta_i} = \frac{\boldsymbol{exp}\left(\boldsymbol{g}(\boldsymbol{h_i})\right)}{\sum_{j=1}^{T} \boldsymbol{exp}\left(\boldsymbol{g}(\boldsymbol{h_j})\right)} \quad \textbf{(4)}$$

where g is a function that maps each hidden state to a scalar, a single-layer feed-forward neural network. The output of the attention layer is the weighted sum of the input sequence, using the attention weights as the weights:

$$\boldsymbol{c} = \sum_{t=1}^{T} \boldsymbol{\alpha_t h_t} \quad \textbf{(5)}$$

The vector c is then concatenated with the current hidden state of the network and passed through a linear layer and non-linear activation function to compute the final output of the network.

The self-attention mechanism is a type of attention mechanism that enables a model to focus on different parts of the input sequence and weigh their importance when making predictions. The input sequence is transformed into three vectors: the query vector, the key vector, and the value vector. The query vector is used to calculate the similarity between the input sequence and a particular element, the key vector is used to store the representation of the input sequence, and the value vector is used to store the output. The self-attention weight for position i is computed as:

$$\alpha_i = \frac{exp(g(h_i))}{\sum_{j=1}^{T} exp\left(g(h_j)\right)} \quad (6)$$

Where h_i is the hidden state of the input sequence at position i, and g is a function that computes the score between the query and the key. The context vector for position i is then computed as:

$$c_i = \sum_{j=1}^{T} \alpha_j h_j \quad (7)$$

This is the same as the equation for global attention, but with the self-attention weights instead of the alignment weights. The context vector *c* is then used as input to the feedforward network (Cao *et al.*, 2021).

3.4 AT-LSTM

Long-short-term memory provides gates in their basic unit. These gates can capture long-term and short-term memory along the time steps and avoid the gradient exploding and vanishing in traditional RNNs. The gates are named forget gate, input gate, and output gate. The LSTM has two memory cells: the candidate memory cell $\widehat{C_t}$, and the memory cell C_t. The memory cell C_t was proposed as a long-term memory to aggregate the relevant information throughout the time steps. The detailed structure of LSTM is presented in Figure 1. The equations for the forward pass of the LSTM cell given a time step *t* are as follows (Freire *et al.*, 2022):

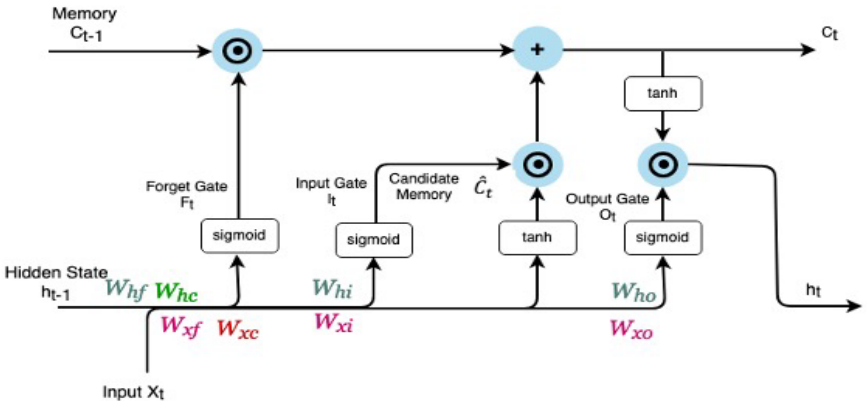


Figure 1: The structure of one LSTM neuron.

$$I_t = \sigma\left(X_t W_{xi} + h_{t-1} W_{hi} + b_i\right) \text{ (8)}$$

$$F_t = \sigma\left(X_t W_{xf} + h_{t-1} W_{hf} + b_f\right) \text{ (9)}$$

$$O_t = \sigma\left(X_t W_{xo} + h_{t-1} W_{ho} + b_o\right) \text{ (10)}$$

$$\widehat{C_t} = tanh\left(X_t W_{xc} + h_{t-1} W_{hc} + b_c\right) \text{ (11)}$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \widehat{C_t} \text{ (12)}$$

$$h_t = O_t \odot tan\, h(C_t) \text{ (13)}$$

$$\hat{y}_t = f(W_h h_t + b_h) \text{ (14)}$$

Where I_t, F_t, O_t, C_t, h_t represent the vector value of the LSTM node at time t in the input gate, the forget gate, the output gate, the cell state, and the output of the LSTM block respectively. W_{xi}, W_{xf}, W_{xo} are the weights between the input layer and the gates at a timestep t. W_{hf}, W_{hi}, W_{ho} are the weights between the hidden recurrent layer and the gates. W_{ci}, W_{cf}, W_{co} are the weights between the cell state and the gates. b_i, b_f, b_o are the biases of the input gate, the forget gate, and the output gate. The output, weights, bias of the activation function f are represented by $\hat{y}, W_h, b_h$,respectively. The set of activation functions consists of the sigmoid function (σ), the element-wise multiplication ($\odot$) and the hyperbolic activation function tanh.

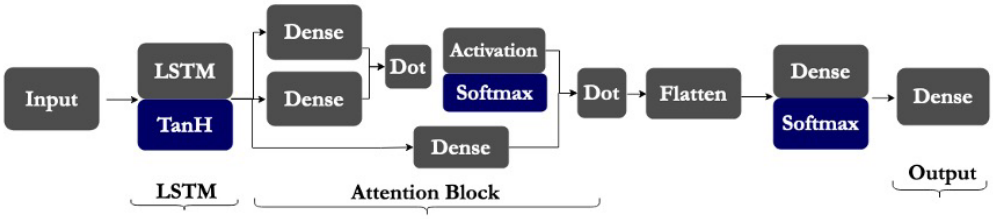


Figure 2: AT-LSTM Model.

Figure 2 shows the structure of the AT-LSTM model. The input to the model is a pre-processed sequence of network traffic records in the form of an LSTM layer. The LSTM layer is followed by a self-attention mechanism, where the query, key, and value are extracted using dense layers.. The attention weights are computed using the dot product of the query and key, followed by a softmax activation function. The attention vector is then computed as the weighted sum of the values using the attention weights. The output of the attention layer is then flattened and passed through a dense layer with a softmax activation function to make the final prediction.

3.5 AT-BiLSTM

Bidirectional LSTM (BiLSTM) uses two separate hidden LSTMs in each layer, one processing the input sequence forward and the other processing the input sequence backwards. The outputs from both LSTMs are concatenated and used as the final hidden representation of the input sequence. To implement BiLSTM, the following changes are implemented to the LSTM equations:

$$\boldsymbol{I'_t = \sigma(X_t W'_{xi} + h'_{t+1} W'_{hi} + b'_i)} \textbf{ (15)}$$

$$\boldsymbol{F'_t = \sigma\left(X_t W'_{xf} + h'_{t+1} W'_{hf} + b'_f\right)} \textbf{ (16)}$$

$$\boldsymbol{O'_t = \sigma(X_t W'_{xo} + h'_{t+1} W'_{ho} + b'_o)} \textbf{ (17)}$$

$$\boldsymbol{\widehat{C_t}' = tanh(X_t W'_{xc} + h_{t+1} W'_{hc} + b'_c)} \textbf{ (18)}$$

$$\boldsymbol{h'_t = tanh(C'_t) \odot O'_t} \textbf{ (19)}$$

$$\boldsymbol{\hat{y}_t = \sigma(h_t W_h + b_h) + \sigma(h'_t W'_h + b'_h)} \textbf{ (20)}$$

For the input gate, forget gate, and output gate equations, new weight matrices, and bias vectors are added for the backward LSTM, denoted by $W'_{xi}, W'_{xf}, W'_{xo}, W'_{hi}, W'_{hf}, W'_{ho}, b'_i, b'_f, b'_o$. For the hidden state update equation, new weight matrices and bias vectors are added for the backward LSTM, denoted by W'_{xc} and W'_{hc}, and b'_c. Then the forward and backward hidden states, h_t and h'_t, are concatenated at each time step to obtain a combined hidden state, denoted by $h_t^{comb} = [h_t, h'_t]$. The output layer equation for each time step would depend on the combined hidden state h_t^{comb}. The gradient updates for the backward LSTM would need to be accumulated separately from the forward LSTM, and the backward LSTM would need to be run in reverse order.

4. Experimental Results

4.1 Setup Environment

The experiment was carried out on Google Jupyter notebooks using the GPU for model training. The model was trained using varied batch sizes and epochs. To prevent overfitting and save computational resources, the model was also set up with an early stopping of a patience of 8 and a minimum delta of 0.001. Table 1 shows the details of the environment specification.

Table 1: Environment Specification

Unit	Description
Processor	Intel(R) Xeon(R) CPU @ 2.30GHz
GPU	Tesla P100-PCIE-16GB

Unit	Description
System RAM	13 GB
OS	Linux 5.10.147+

4.2 Results and Discussions

Accuracy, precision, recall, and F1 score are used for assessing the efficacy of the model. Figure 3 and Figure 4 show confusion matrices, which are used to measure how well classification models work. The results of the models are visualised by comparing the actual classes of attacks to the predicted attack classes made by the models. The diagonals of the matrix show where the actual class and the predicted class match up. The off-diagonal values represent the cases where the model has made incorrect predictions.

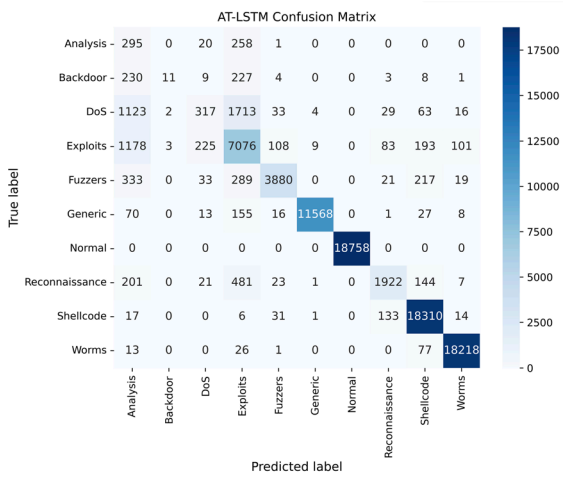


Figure 3: AT-LSTM confusion matrix

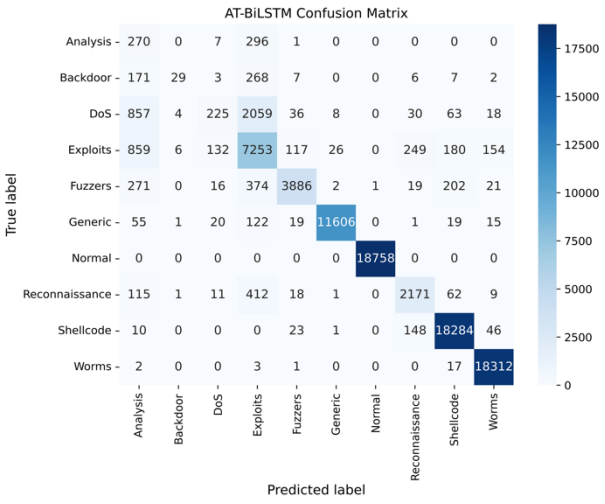


Figure 4: AT-BiLSTM confusion matrix

Table 2 displays the performance metrics of two intrusion detection models, AT-LSTM and AT-BiLSTM, on the dataset. The models were evaluated based on their accuracy, precision, recall, and F1-score. Both models have a high accuracy, with AT-LSTM having an accuracy of 89.90% and AT-BiLSTM having an accuracy of 92.20%. The precision of both models is also high, with AT-LSTM having a precision of 93.06% and AT-BiLSTM having a precision of 93.83%. This suggests that both models can correctly identify most positive instances. The recall of the AT-LSTM model is 87.07%, while the recall of the AT-BiLSTM model is 91.19%. The F1-score values for the AT-LSTM and AT-BiLSTM models are 89.96% and 92.50%, respectively. A higher F1-score value indicates a better balance between precision and recall, which implies better overall performance of the model. In this case, the AT-BiLSTM model has a slightly higher F1-score value, indicating that it may be a better model than the AT-LSTM in terms of overall performance. BiLSTM model is advantageous due to its ability to process the input sequence

in both forward and backward directions. This allows the model to capture more contextual information and dependencies within the sequence, which is particularly useful for detecting network intrusions. Attacks may involve complex and subtle patterns of activity that occur over time, that may not be evident from a single direction.

Table 2: Performance Metrics of AT-LSTM and AT-BiLSTM

Model	Performance Metrics			
	Accuracy	Precision	Recall	F1-Score
AT-LSTM	89.90%	93.06%	87.07%	89.96%
AT-BiLSTM	92.20%	93.83%	91.19 %	92.50 %

The AT-LSTM and AT-BiLSTM deep learning models were evaluated on various attack categories to assess their effectiveness in detecting network intrusions. The results in Table 3 demonstrate that both models have high detection rates for most of the attack categories. Notably, the models achieved a 100% detection rate for generic and normal network flows, which are the most common types of network traffic. This indicates that the models are highly effective in identifying normal network activity. For attacks that involve large-scale network scans and automated propagation, such as worms and fuzzers, the models achieved high detection rates. The AT-LSTM model had a detection rate of 89% for fuzzers, while the AT-BiLSTM model achieved a higher rate of 92%. Similarly, the AT-LSTM model detected 97% of worm attacks, and the AT-BiLSTM model achieved a detection rate of 98%. These high detection rates demonstrate the models' effectiveness in detecting attacks propagating quickly and widely across a network. The models also performed reasonably well in detecting reconnaissance, backdoor, exploits, and shellcode attacks. For example, the AT-LSTM model achieved a detection rate of 75% for reconnaissance, while the AT-BiLSTM model achieved a higher rate of 85%. The AT-LSTM model detected 79% of backdoor attacks, while the AT-BiLSTM model achieved a higher rate of 100%. In addition, the AT-LSTM model achieved a detection rate of 78% for exploits, while the AT-BiLSTM model detected 73%. Finally, the AT-LSTM model detected 93% of shellcode attacks, while the AT-BiLSTM model achieved a slightly higher rate of 96%. These attack categories can be particularly challenging to detect due to their sophisticated techniques that can evade traditional security measures. However, the high detection rates of the models for these attacks demonstrate their effectiveness in identifying such threats. For instance, the backdoor attack is a type of malware that allows attackers to access a system remotely and control it without the user's knowledge. The high detection rate of 100% achieved by the AT-BiLSTM model for this attack type indicates its ability to detect such subtle and complex activity patterns.

Table 3: Detection Rate of Attacks Categories

Attack Type	Detection Rate	
	AT-LSTM	AT-BiLSTM
Analysis	7%	8%
Backdoor	79%	63%
DoS	43%	51%
Exploits	78%	73%
Fuzzers	89%	92%
Generic	100%	100%
Normal	100%	100%
Reconnaissance	75%	85%
Shellcode	93%	96%
Worms	97%	98%

The accuracy vs. epochs plot is a keyway to measure how well the models are doing while they are being trained. It is used to figure out if the model is too good or too bad and to find the best number of epochs. Figures 4 and 5 show that the two models' training accuracy steadily improves with each epoch. This means that the models are learning from the data and getting better over time. The validation accuracy, which assesses the model's ability to perform well on new, unseen data, exhibits some fluctuations, particularly in the AT-BiLSTM model,

despite an overall upward trend. This may be due to overfitting, where the complex AT-BiLSTM model is too closely fitted to the training data, resulting in variability in its performance on new data. As a result, the training accuracy may continue to rise with each epoch, while the testing accuracy may fluctuate.

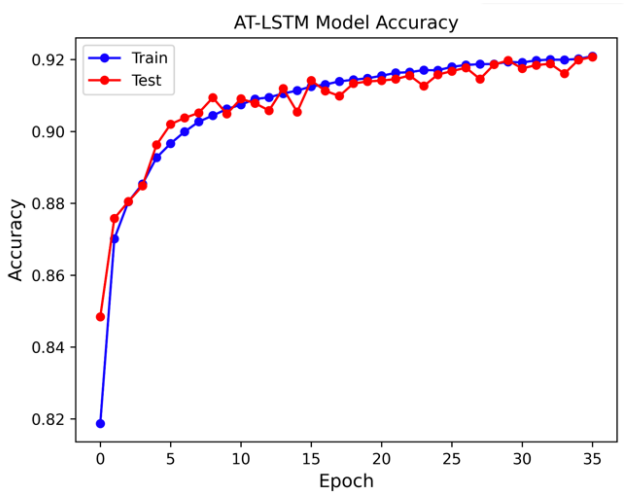


Figure 4: AT-LSTM Accuracy vs. Epochs.

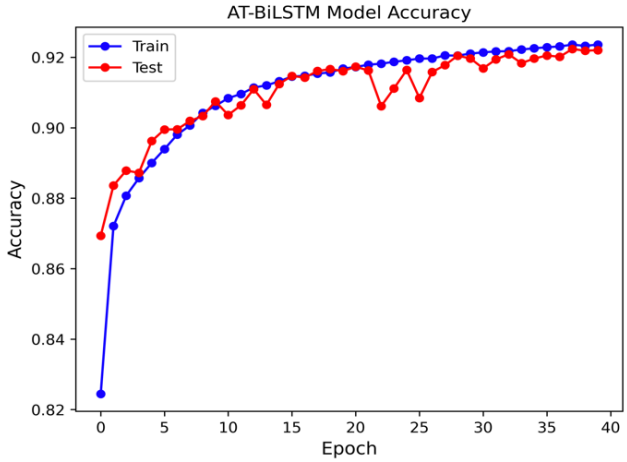


Figure 5: AT-BiLSTM Accuracy vs. Epochs.

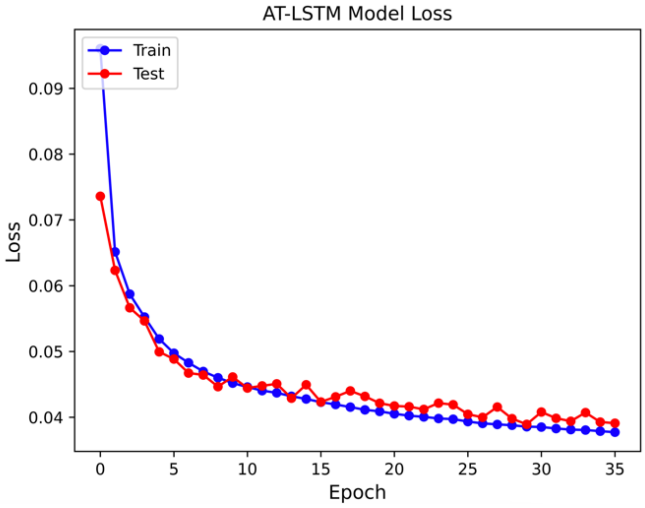


Figure 6: AT-LSTM Loss vs Epochs.

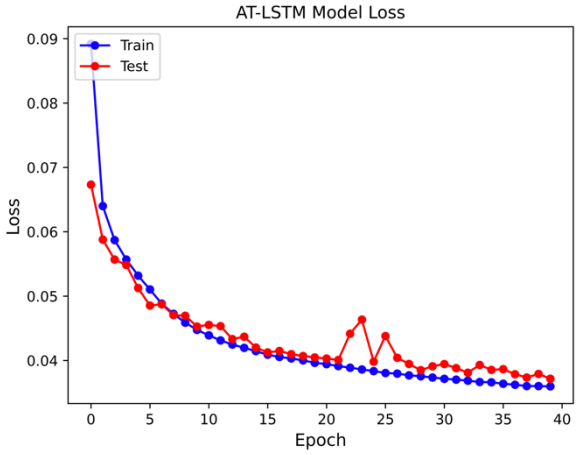


Figure 7: AT BiLSTM Loss vs Epochs.

The loss vs. epochs plot shows how the training and validation loss values change as the model trains over a certain number of epochs. The loss is a measure of how well the models can predict the target output. In Figures 6 and 7, we can see that as the number of epochs goes up, both the training loss and the validation loss tend to go down. Figure 8 shows that the AT-LSTM model with smaller batch sizes tends to converge faster and achieve higher accuracy than those with larger batch sizes. This is because smaller batch sizes allow for more frequent parameter updates, leading to faster convergence and better generalisation. On the other hand, larger batch sizes may converge more slowly and suffer from overfitting due to seeing a biased representation of the overall dataset. Thus, the figure shows a clear trend of increasing accuracy with smaller batch sizes over the course of training.

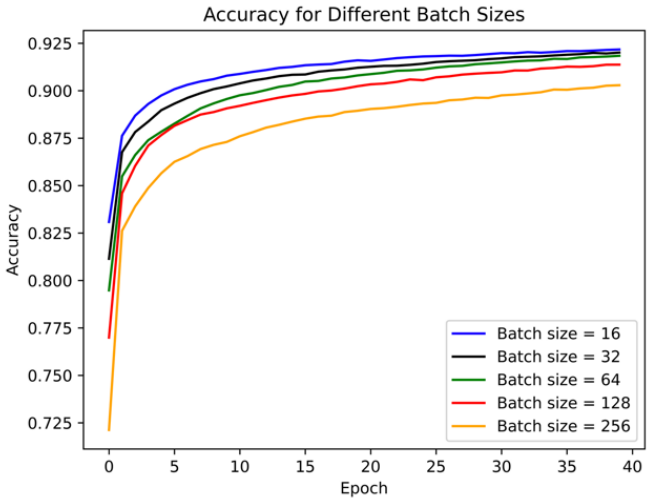


Figure 8: Accuracy vs Epochs for different batch sizes.

5. Conclusion

The study demonstrated the effectiveness of attention-based LSTM and BiLSTM models in detecting network intrusions with precision rates exceeding 93%. The models could capture complex and subtle patterns of activity associated with different types of attacks. However, to ensure their practical applicability, the models' limitations must be addressed in future research. This includes developing models that are more sensitive to the characteristics of DoS attacks, such as high traffic volume and unusual traffic patterns, and incorporating data from multiple sources to provide a more comprehensive view of network activity. Additionally, incorporating information about the source of the traffic, such as IP addresses, could help to identify better and block malicious traffic sources. Finally, developing methods to identify and mitigate adversarial attacks, such as adding noise to the input data or using robust training techniques, could further enhance the security and effectiveness of the intrusion detection system.

Acknowledgement

This work was supported by the United Arab Emirates (UAE) University Program for Advanced Research (UPAR), under Grant number 31T122.

References

Aldallal, A. (2022) 'Toward Efficient Intrusion Detection System Using Hybrid Deep Learning Approach', *Symmetry*, 14(9), p. 1916.

Bakhsh, S.T. *et al.* (2019) 'An adaptive intrusion detection and prevention system for Internet of Things', *International Journal of Distributed Sensor Networks*, 15(11), p. 155014771988810.

Benaddi, H. *et al.* (2022) 'Robust Enhancement of Intrusion Detection Systems Using Deep Reinforcement Learning and Stochastic Game', *IEEE Transactions on Vehicular Technology*, 71(10), pp. 11089–11102.

Cao, K. *et al.* (2021) 'Network Intrusion Detection based on Dense Dilated Convolutions and Attention Mechanism', in *2021 International Wireless Communications and Mobile Computing (IWCMC)*. *2021 International Wireless Communications and Mobile Computing (IWCMC)*, Harbin City, China: IEEE, pp. 463–468.

Fan, S. *et al.* (2019) 'ALEAP: Attention-based LSTM with Event Embedding for Attack Projection', in *2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC)*. *2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC)*, London, United Kingdom: IEEE, pp. 1–8.

Freire, P.J. *et al.* (2022) 'Computational Complexity Evaluation of Neural Network Applications in Signal Processing'. arXiv.

Gaifulina, D. and Kotenko, I. (2021) 'Selection of Deep Neural Network Models for IoT Anomaly Detection Experiments', in *2021 29th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*., Valladolid, Spain: IEEE, pp. 260–265.

Han, X. *et al.* (2020) 'STIDM: A Spatial and Temporal Aware Intrusion Detection Model', in *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. *2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, Guangzhou, China: IEEE, pp. 370–377.

Laghrissi, F. *et al.* (2021) 'IDS-attention: an efficient algorithm for intrusion detection systems using attention mechanism', *Journal of Big Data*, 8(1), p. 149.

Lan, M. *et al.* (2020) 'A Novel Industrial Intrusion Detection Method based on Threshold-optimized CNN-BiLSTM-Attention using ROC Curve', in *2020 39th Chinese Control Conference (CCC)*. *2020 39th Chinese Control Conference (CCC)*, Shenyang, China: IEEE, pp. 7384–7389.

Liu, C., Gu, Z. and Wang, J. (2021) 'A Hybrid Intrusion Detection System Based on Scalable K-Means+ Random Forest and Deep Learning', *IEEE Access*, 9, pp. 75729–75740.

Liu, D. *et al.* (2020) 'Air pollution forecasting based on attention-based LSTM neural network and ensemble learning', *Expert Systems*, 37(3).

Liu, Q. *et al.* (2022) 'Improving wireless indoor non-intrusive load disaggregation using attention-based deep learning networks', *Physical Communication*, 51, p. 101584.

Mighan, S.N. and Kahani, M. (2021) 'A novel scalable intrusion detection system based on deep learning', *International Journal of Information Security*, 20(3), pp. 387–403.

Mirza, A.H. and Cosan, S. (2018) 'Computer network intrusion detection using sequential LSTM Neural Networks autoencoders', in *2018 26th Signal Processing and Communications Applications Conference (SIU)*. *2018 26th Signal Processing and Communications Applications Conference (SIU)*, Izmir: IEEE, pp. 1–4.

Moustafa, N. and Slay, J. (2016) 'The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set', *Information Security Journal: A Global Perspective*, 25(1–3), pp. 18–31.

Sarhan, M. *et al.* (2021) 'NetFlow Datasets for Machine Learning-Based Network Intrusion Detection Systems', in Z. Deze et al. (eds) *Big Data Technologies and Applications*. Cham: Springer International Publishing (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering), pp. 117–135.

Saurabh, K. *et al.* (2022) 'LBDMIDS: LSTM Based Deep Learning Model for Intrusion Detection Systems for IoT Networks'. arXiv.

Waqar, M. *et al.* (2021) 'An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction', *Scientific Programming*. Edited by M.-C. Chen, 2021, pp. 1–12.

Wei, P. *et al.* (2023) 'A novel intrusion detection model for the CAN bus packet of in-vehicle network based on attention mechanism and autoencoder', *Digital Communications and Networks*, 9(1), pp. 14–21.

Wu, Z. *et al.* (2022) 'RTIDS: A Robust Transformer-Based Approach for Intrusion Detection System', *IEEE Access*, 10, pp. 64375–64387.

Xu, C. *et al.* (2018) 'An Intrusion Detection System Using a Deep Neural Network With Gated Recurrent Units', *IEEE Access*, 6, pp. 48697–48707.

Yan, Y. *et al.* (2020) 'A Network Intrusion Detection Method Based on Stacked Autoencoder and LSTM', in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*. *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, Dublin, Ireland: IEEE, pp. 1–6.

Yang, H. *et al.* (2022) 'Intrusion Detection Model For Power Information Network Based On Multi-layer Attention Mechanism', in *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*. *2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, Chongqing, China: IEEE, pp. 825–828.

Yu, Y. *et al.* (2022) 'LSTM-Based Intrusion Detection System for VANETs: A Time Series Classification Approach to False Message Detection', *IEEE Transactions on Intelligent Transportation Systems*, 23(12), pp. 23906–23918.

Zhang, C. *et al.* (2022) 'Comparative research on network intrusion detection methods based on machine learning', *Computers & Security*, 121, p. 102861.

Zhang, Y. *et al.* (2018) 'Deep Learning Intrusion Detection Model Based on Optimized Imbalanced Network Data', in *2018 IEEE 18th International Conference on Communication Technology (ICCT)*. *2018 IEEE 18th International Conference on Communication Technology (ICCT)*, Chongqing: IEEE, pp. 1128–1132.

Zhang, Y. *et al.* (2019) 'A Text Sentiment Classification Modeling Method Based on Coordinated CNN-LSTM-Attention Model', *Chinese Journal of Electronics*, 28(1), pp. 120–126.

Towards the Development of Indicators of Fake Websites for Digital Investigation

Mera Alremeithi [1], Aysha Alkuwaiti [2], Haya Alobeidli[3] and Richard Ikuesan
Computing and Applied Technology, College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates

meralremeithi@gmail.com
ayshaalkuwaiti5@gmail.com
haya.alobeidli01@gmail.com
richard.ikuesan@zu.ac.ae

Abstract: A fake website is considered a website that is intended to cause harm and manipulate users, especially novice users without some knowledge of indicators of fakeness. Understanding the indicators of fake websites is thus considered an important concept to avoid being a victim of malicious attacks in online engagements. In some cases, such knowledge is required to reduce the potential attack surface of cyber criminals. However, the increasing rate of website diversity and complexities makes it difficult for an individual to distinguish between a fake and a real website while compounding the investigation process of a website. Also, the growing rate of website imitation technology and website domain closure presents a veritable platform for the development of fake websites. As a step towards determining the genuineness of a website, this study developed a forensic framework based on an exploratory analysis of different genres of fake websites. To achieve this, forensic methodologies and processes were applied to methodically selected samples of known fake websites based on three fakeness categories: Hoaxes, Cybersquatting, and Sweepstakes. The result revealed the existence of salient markers which can be used as indicators of fakeness and can be applied across a wide genre of websites. Furthermore, the resultant observation was used to develop a digital forensic framework for website fakeness evaluation. The developed framework was benchmarked to the ISO 27043/2015 and the NIST SP800-86 standard for completeness and relevance to forensic investigation processes. By leveraging the proposed digital forensic framework, an investigation can develop a reliable pointer to evaluate the genuineness of any website, which can significantly reduce the investigation time. For a non-forensic individual, the developed framework can be leveraged to identify, at first glance, the degree of fakeness of a website. Such a mechanism can therefore provide a useful tool to reduce the potential susceptibility of users thereby creating user awareness.

Keywords: Fake website, Hoax, Cybersquatting, Sweepstakes, Fake website indicators.

1. Introduction

There is a growing trend toward attacks that take advantage of the information present online to target human weaknesses. In electronic markets, fraud is a common challenge, and it affects various Internet consumers (Chua & Wareham, 2004). Asa significant source of online fraud, fake websites now generate billions of dollars in false money at the cost of unwary Internet users (Yue Zhang et al., 2007). According to a recent survey, fraudulent websites make up about 20% of the total Web (Gyongyi & Garcia-Molina, 2005). A random sample of more than 105 million web pages found that 35% of ".us" domain pages and 70% of ".biz" domain pages were fake (Ntoulas et al., 2006). Fake websites not only cause immediate financial losses but also have long-term trust-related effects on users that may make them reluctant to do further online transactions (Malhotra et al., 2004). Several classes of fake websites have been identified in the literature. Ranging from a phishing-induced website that seeks to exploit unsuspecting entities to hoaxes for disinformation (Abbasi & Chen, 2009; Chan, 2022; Frischknecht, 2021; Tacchini et al., 2017; Zahedi et al., 2015). A synopsis of the commonly encountered fake website is further presented in Table 1. The approaches leveraged to craft these websites can range from simplistic techniques to a multimodal complex approach.

Table 1: Summary of Techniques used in commonly encountered fake websites.

Techniques/types	Description
Hoaxes	Creation of fake information websites for disinformation and other malicious act to deceive (or sway decisions in some instances) unsuspecting entities.
Cybersquatting	Cybersquatting requires the registration of domains that resembles existing popular/common trademarks with malicious intent. Examples include typosquatting, Homograph-squatting, Combo-squatting, Top-Level Domain-squatting
Domain spoofing	Also referred to as domain masquerading, and similar to typosquatting, is a form of fake website that involves the use of a similar domain (or email) address to an existing domain or email address to deceive an unsuspecting entity
Content spoofing	This technique, also called content injection, involves the injection of arbitrary content into a website to lure unsuspecting entities.

Sweepstakes	This technique leverages the unsuspecting entity's desire to win a lottery or some prized item. In this technique, an attacker could present a pop-up ad on the website to attract the potential of some prize and rewards for a given action. This action would ideally be genuine, however, used for malicious intent.
Scareware scam	This technique also leverages pop-up ads, however, with a psychological component that attempt to scare the potential victim to take an action. Such Ads would suddenly appear on a website with a dire warning and consequential non-committal thus, prompting the potential victim to act. This technique also leverages. Using this technique, other forms of scams be carried such as fake anti-virus software pop-up windows can be conducted.
Phishing website scam	This technique involves the use of a false domain that is similar in name and identity to an official website, thereby luring an unsuspecting entity into an illegitimate site. This technique often redirects the unsuspecting entity to the legitimate site after harnessing the login credential of such an entity. This can also be termed as a Counterfeit of a goods site. Attackers could further leverage this technique to conduct online shopping scams, Charity fraud scams, as well as health website scams.
COVID-19 Vaccine Trial Scam websites	This technique leverages the need for Covid-19 related information ranging from medical supplies, vaccines, and other medical-related equipment. This often involves the creation of a fake website with relevant details (often curated from legitimate websites) albeit, with malicious intent to unsuspecting entities. This technique can be used to craft several phishing scams as well as domain spoofing.
Advertising scam: Astroturfing	This technique utilizes a manipulative behavioral approach to artificially create an impression or disinformation to unsuspecting entities. The combination of this technique and other forms of phishing techniques can be used to generate a more complex fake website with a different audience as the target. A classic example of such is the Stock market scam where fraudulent entities present fake information and statistics about a given commodity to engineer fear and drift in market capitalization. Other forms of Stock market scams could include the development of a temporary website with fake information and rumor targeted at creating chaos, biased trading, and panic buying within a targeted market group.
Concocted website	This technique is used to develop fake sites that appear as legitimate commercial entities, which often pose as real escrow, financial, delivery, or retail companies.
Fake escrow websites	an escrow service is intended to serve as a trusted third-party protection against Internet fraud. A fake escrow website, therefore, acts in the same capacity, however, with a deviation from the standard of trust.

Numerous solutions have been proposed for automatic fraudulent website detection in response to these worries (Li & Helenius, 2007). The majority are lookup tools that only use blacklists made up of unified resource locators (URLs) culled from member-reporting repositories kept by online trading forums. These systems are reactive by nature since they rely on user reports; by the time fake websites are added to the blacklist, many users have already encountered them (Neil Chou et al., 2004). A related family of systems employs proactive classification methods that can identify fraudulent websites without the aid of user reports. These systems make use of fraud signals, which are crucial design components of fake websites that could reveal their falsity. Unfortunately, contemporary classifier systems use extremely simplistic fraud triggers and classification assumptions, making them simple to evade (Li & Helenius, 2007). The detection of bogus websites is also a dynamic issue. Fraudsters frequently deploy new tactics and make use of more advanced technologies (Dinev, 2006). Systems for detecting fake websites have not been able to keep up with improvements made by its competitors. As a result, these technologies are not very good at identifying bogus websites, with most instances seeing detection rates below 70% (Yue Zhang et al., 2007). The types of fake websites that can be detected by existing methods are, thus, likewise constrained (Abbasi & Chen, 2013).

There is, therefore, still a demand for fake website detection tools that can recognize different types of fake websites effectively. This study explored the fakeness of websites from the forensic perspective with a core emphasis on recurring patterns in the different classes of fakeness. Given the numerous numbers of potential fake websites, this study further developed a selection criterion based on the aim of the study**.** Three types of fake websites were selected, which include: Fakeness due to hoaxes, cybersquatting, and sweepstakes.

2. Methodology

The overall process model followed to achieve this study is presented in Figure 1. It comprises three interconnected phases. The first phase presents the process of selecting a fake website genre. To enhance the forensic process of fake website detection, this study adopted the knowledge of inclusion and exclusion criteria

commonly used in design science research. To achieve this, carefully crafted inclusion and exclusion criteria for fake website exploration were developed, as presented in Table 2. By applying these criteria to the identified techniques presented in Table 1, this study further reduced the explored fake website techniques to be studied to three. To apply the selection criteria, all techniques identified in Table 1 met these inclusion criteria. Criterion IC6 was specifically considered for all the techniques. However, when the exclusion criteria were applied, the list was reduced to three. These include Hoaxes, cybersquatting, and sweepstakes. The selected techniques cover the general techniques for fake website development.

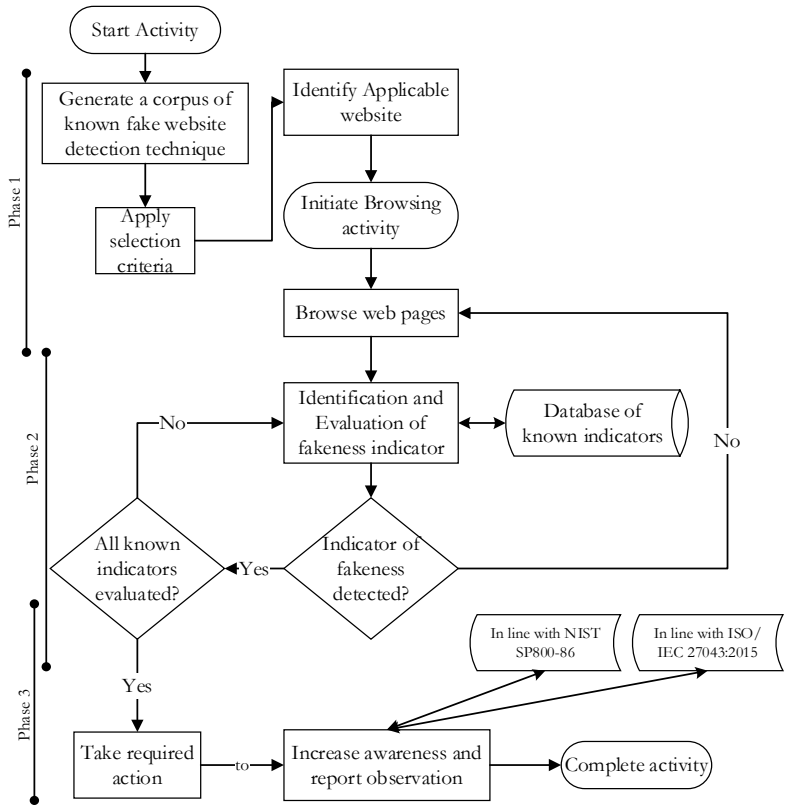


Figure 1: Operational framework for an indicator of fakeness detection

Table 2: Selection Criteria for explored fake websites.

Inclusion criteria	Exclusion criteria
IC1: Potential to trigger a further attack or be combined to craft a more complex attack	EC1: Has been extensively studied in the literature
IC2: Performed with minimal effort and simplicity	EC2: Requires complex mathematical formation and modeling such as the use of machine learning and deep learning algorithms to detect
IC3: Pervasive and relatively common	EC3: Requires the use of a blacklist of either IP, URL, or other such properties for detection
IC4: Difficult to detect by an unsuspecting entity	EC4: Focuses on social media news only.
IC5: Prevailing technique despite an extensive awareness campaign	EC5: Potentially volatile such that relevant indicators may not available after the fact.
IC6: Focuses on websites.	

In the second phase, an initial process a user may undertake is browsing the web. Primarily, the user might not recognize that the browsed website is a counterfeit. This phase of the process model provides a mechanism for evaluating the website for fakeness, by carefully comparing legitimate and fake websites for potential indicators of fakeness. By leveraging the forensic processes, an indicator of the fakeness database is further built, iteratively. The third phase considers the process of remediating fakeness exploitation. If a user discovers indicators on the website, they must ensure that all the indicators are studied and evaluated using the knowledge base. If all indicators are evaluated and any indicators match with the website, then the user is required to remediate through an awareness program. This approach can be used to prevent other users from

becoming victims of the website and report the detection of the indicators. However, if no indicators have been discovered, the user can continue browsing the web. Using this process model, as demonstrated in Figure 1, this study further conducts case studies with results presented in the subsequent sections. Primarily, the processes are evaluated in compliance with NIST SP and ISO/IEC standards.

3. Result and Analysis

Phases 2 and 3 of the process model presented in Figure 1 are applied to the three genres of fakeness in this section.

3.1 Understanding Indicators of Fakeness in Hoaxes

Without a doubt, it is well known that currently, we live in a fast-paced world where various events occur, and different news are spread online regarding these events. This caused misinformation to be formed and drastically spread among people digitally. When individuals are exposed to false information, it may lead them to perform actions that can cause them extreme damage. The terminology that describes this form of information is referred to as Hoaxes. A hoax can spread in different forms including fake information, virus messages, chain letters, or the spread of fake news websites. Hoax websites are created for reasons such as intentionally misleading the viewer, injecting malware, for political or financial gain, or spreading false information.

3.1.1 Indicators of Hoaxes on Websites

a. Clear Signals:

An obvious signal is that if the user is prompted to a site that requests to forward the statement to various people, it is clear it is meant to spread false information to many people. A hoax can be identified when it does not contain any credibility of the source information. Also, if a suspicious URL is included in the message, it is a direct identifier that it is a hoax since it could prompt the user to a fake website. Moreover, some statements are cleverly composed to grab the viewers' attention and provoke them emotionally to make the message seem believable.

b. Impossible Statements and User Manipulation:

A hoax aims to cause the recipient to be shocked by a warning that seems impossible to occur. Statements that seem impossible to believe are often spotted quickly. An example of an unreasonable message could be "A virus is detected, and it will cause fire and your computer will explode". Attackers usually create those alerts and make it seem that it is coming from an official organization.

c. Fact-checking:

In some cases, when a user is prompted to message or article that seems to be suspicious, an extensive fact-checking procedure needs to be implemented. It is recommended to check the source. For instance, the web address of the page should be inspected carefully because there might be spelling errors in the URL or domain name extensions. The authority of the page should be examined to determine if the mentioned author is real and credible.

3.1.2 Case Study-1:

During the spread of COVID-19 in 2020, cyber criminals exploited that period to create scams which led many people to be misled by those scams. Multiple organizations have created websites to display and track any statistics on COVID-19. However, cybercriminals came up with a method to take advantage of the coronavirus map websites. Attackers were able to design websites similar to the official pandemic maps that display statistics of the event. (Tangermann, 2020). The creation of such websites caused an increase in spreading false information regarding the virus and allowed hackers to obtain users' information. According to a security researcher, Shai Alfasi, attackers utilized these maps to discover the user's data (Mehta, 2020). The main malicious approach implemented by attackers was creating the coronavirus map websites that prompt the user to install an application to track any updates of the event. It was purposely designed this way by hackers to create a harmful binary file and download it to the user's system. The researcher has reported that hackers performed this technique by using software called AZORult. The software obtained information from the user's

computer including cryptocurrencies, passwords, installed programs, and stored account information (*Azorult Malware Information*, 2019).

3.1.3 Forensic Perspective

From the concepts discussed so far, it is determined that the main indicator of hoaxes is by performing fact-checking. Therefore, a simple google search about tools or sites to use to identify misinformation would be helpful in this encounter. This seems a direct and clever approach to confirm the authenticity of a certain statement or story, however, not any fact checker site is trustworthy. It is undeniable that the number of misleading statements diffused in the online ecosystem not only confuses the viewer regarding the legitimacy of the information but also has doubts regarding the fact-checking sites that claim to be trustworthy. This brings us to an important aspect that we should be aware of which is that even the fact checkers' tools that seem to be legitimate are in need to be inspected. “Who inspects the fact-checkers?” is a question that is important to discuss when it comes to analyzing a source. Therefore, the presence of digital forensic approaches is required.

From a generic forensic perspective, there are basic approaches that are implemented. The study (Sousa-Silva, 2022) highlighted two generic techniques. The first approach mentioned is “human intervention” which involves human interaction with the data set to be examined. Human fact-checkers are recruited to confirm the integrity of the information. This approach may be useful in some cases; however, it has obvious limitations. Since facts are checked manually, generating results is a slow process since it is time-consuming. There is a high rate of producing inaccurate results because human fatigue is an expected occurrence. The use of this approach depends on the claim, the subject of the target, and how effective the fact-checking method is.

An alternative to the discussed approach includes the use of algorithms to indicate inaccurate content and verify sources (Sousa-Silva, 2022). The algorithm in this case focuses on the content of the data and the rate of content-checking accuracy instead of focusing on the sources’ reputation. Its emphasis is the message’s propagation dynamics in which the dynamics of the theory are being used to recognize the credibility of the information (Baror et al., 2020; Kebande et al., 2020). The main advantage of this technique is that it avoids any bias and can quickly process a humongous volume of data without the need for human interaction. All tools or mechanisms created so far to detect fake news properly, according to the author (Sousa-Silva, 2022), are not considered appropriate in the digital forensic field.

3.2 Understanding Indicators of Fakeness in Cybersquatting

Cybersquatting refers to a situation where a criminal uses, registers, or sells a domain name to benefit from another person’s or organization’s trademark (Mercer, 2000). The fast growth of the practice can be explained by the fact that most companies lack people who have deep technological knowledge. This observation has been evident since 2010 when cybersquatting started to affect the retail industry to a large extent, as demonstrated in the table below. In this case, the criminals' registered domains by using the names of certain popular companies to resell those later (Orduña-Malea, 2021).

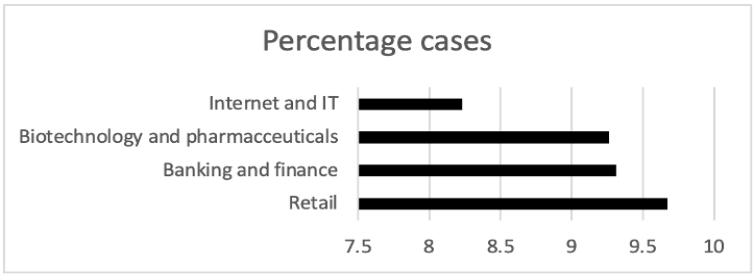


Figure 2: Cybersquatting Effect on Different Industries (Petrosyan, 2022)

3.1.4 Indicators of Cybersquatting on Websites

Several indicators can reveal evidence of cybersquatting. The victims of the practice can identify it by searching for their domain names and looking for websites that mimic their platforms, have many ads but lack content, communicate negative information about their products, or have been under construction for a long duration.

a. Identical URLs:

The first indicator of cybersquatting is whether the desired URL leads to the domain page on sale instead of the unique website. Though this indicator may not independently prove a domain shark, it can point to a possible cybersquatting case, especially a type known as typosquatting (Agten et al., 2015). For example, a cyber-squatter who intends to hijack a site or a URL can simply aim to identify and convert Internet users' common typing mistakes. For instance, the squatters know that a person who intends to type www.buisness.com may type www.bizness.com or www.business.net, which comprise top-level domains. In this instance, the squatters will attempt to take advantage of a visual, logo, content, layout, color, sound, or hardware trademark similarities.

b. *Identical Products:*

The second indicator of cybersquatting is that the user will always be taken to a page that is promoting or selling related products. To achieve results, the criminal chooses a popular brand and associates it with similar products. This practice can benefit them since they use a long-term positive reputation of the products that another company provides. Consider as an example, company X sells medical devices, and website A sells related products and/or services, as well as having similar trademarks, content, and so on, then company X might be enduring a case of cybersquatting.

c. *Domain Name Displayed for Sale:*

Finally, other indicators become visible when the criminals attempt to proactively sell the site and try to overcharge their potential clients. In this case, the criminals are the ones who voice their intention to sell the domain before the owner expresses any interest to buy one (Chandra & Bhatnagar, 2019). The seriousness of squatting is demonstrated in the screenshot below, where cybersquatters are attempting to sell domains of a well-known landmark at an exorbitant price.

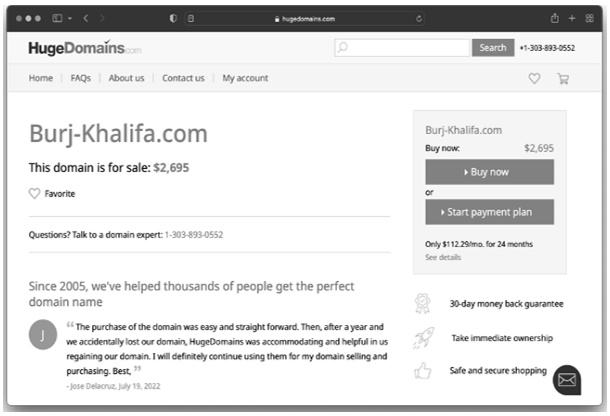


Figure 3: Domain for Sale

3.1.5 Case Studies-2:

Among the common real-life case studies of cybersquatting is the case that pitted Tom Cruise against Jeff Burger in 2006. By the time Cruise learned about the crime, the squatter Burger had been using the domain based on his name (TomCruise.com) for over a decade (*Cybersquatting Examples: Everything You Need to Know*, 2022). A click on the website gives users an option of being redirected to watch the Top Gun Maverick 2022 new trailer. As a result of the *Tom Cruise v Operations Carter* case, the domain was transferred to Tom Cruise where he won the case.

In another case, a person has attempted to purchase domains to impersonate Amul which is one of the largest dairy firms in India, and created phishing sites. Amuldistributor.com, Amulboard.com, Amufran.com, Amufran.org.in, and Amuldistributorindia.com are examples of the sites to which Amul fell as a victim (Agarwala & Kang, 2021). To pull off their scam, the perpetrators created fictitious bank accounts in Amul's name, demanded money as a subscription to become an Amul distributor, and franchise-operated employment scams and demanded payment from applicants for job postings.

3.1.6 Forensic Perspective

In most countries, the forensic tools to use, the specific indicators to focus on, and the application process to follow when investigating cybersquatting are indicated in a domain dispute resolution policy ("Domain Names and Cybersquatting," 2017). For forensic purposes, the authorities search for three major factors. The domain

name in question should be either confusingly similar to the service mark or a trademark of the complainant (Karanicolas, 2020). In addition, the registrant's interest in or right to the domain name should not be legitimate. Finally, the questioned domain name must have been registered and is currently being used for an illegal purpose or in bad faith (Radhika Vivek Bhusari & Karan Ramchandra Rampure, 2022).

Besides, there are multiple forensic tools available for performing a domain investigation, which seeks to determine whether a domain is authentic. Firstly, it is important to note that any domain usually has an IP address as well as host names that can legally send emails (Nangia, 2021). Secondly, it can be found in the Sender Policy Framework (SPF) record. As such, the process of investigating cybersquatting is focused on obtaining the IP address from which an email has been sent; then a verification process begins to confirm if it is like the emails in the SPF record (Singh, 2018). In addition, one of the online tools used in the investigation process is the Domain Name System (DNS), which helps in verifying the similarity between the suspicious and the authentic domain (Hewling, 2013). As a naming system, DNS allows users to reach any website by taking the keyed website name and redirecting the user to the IP address associated with it. The specific DNS tools commonly include DNSlytics, DomainEye, and Domain Dossier. For instance, the Domain Dossier tool creates reports from publicly available information on domain names and IP addresses to assist in problem-solving and cybercrime investigation. These reports could demonstrate the contact details for the domain owner, registry, and registrar details, as well as the organization responsible for hosting the website.

3.3 Understanding Indicators of Fakeness in Sweepstakes

Sweepstake scams are going widely distributed for a long time now, and yet, still they are going on strong. The initial contact in a sweepstake might be a notification you get while browsing a website that offers congratulations for winning a big contest. You will be asked to provide your bank account details. Once they have someone in their clutches, they will continue to contact them promising the big prize is just one more payment away. Most of the time common sense wins out and the customer recognizes that this is just another online scam. However, a less experienced user could not realize the threat until it is too late.

Figure 4: A pop-up Sweepstake

As shown in Figure 4, the user is tempted by the "Congratulations You Are Today's Winner" message. It's the ideal method for phishers to draw visitors into their web of deception.

3.1.7 Indicators of Sweepstakes on Websites

A sweepstakes win would be a dream come true. However, if what you initially believe to be a genuine win notification turns out to be a sweepstakes fraud, your dream could become a nightmare.

a. Sweepstakes demand payment to receive a prize:

Scammers may claim that you must pay for the following before they will deliver your prize: Taxes on sweepstakes, Customs fees, Shipping fees, and service charges. If that's the case, you're probably dealing with a con. You will never be required to pay a charge to receive a reward in legitimate sweepstakes.

b. The fact that you won a contest you didn't enter:

You can only win sweepstakes that you have entered. It's a warning sign if you get a win notification from a contest that you don't recall entering. Take the time to conduct further research before you answer.

c. Pressuring you to act in a hurry:

Are you under any time constraints to reply to your win notification? If so, move cautiously. The reason why sweepstakes scammers want you to act quickly is so they can have your money before your check bounces.

d. Asks you for your bank information:

Does receiving your reward require you to validate your bank account information? This is blatant evidence of sweepstakes fraud. Legitimate sweepstakes don't require your credit card number for information verification. Your social security number is the only private data a trustworthy sweepstakes sponsor requires to handle your win.

e. Sweepstakes scams don't know your first and last name:

Does the title "Dear Winner" appear in your win notification? If so, this is a serious red flag. Numerous sweepstakes frauds send many letters or emails to every address they can find, without knowing any personal information. Genuine sweepstakes already have your entry information from the entry form.

3.1.8 Case Studies-3:

A crime was reported by GULF news in 2017 about a victim who clicked on a pop-up notification from a certain website as they claim he was the winner of 500,000 DH "in a lucky draw". He has been told to receive his prize; they should get his card details number including his pin. The victim proceeded further by giving them all the details about his bank account. After a while, he noticed that a withdrawal procedure was made from his card to an unknown person. At this moment, the victim realized that he was trapped in a fake prize situation. Directly he contacted Abu Dhabi police and by coordination with authorities, they managed to refund the victim's money back (*Police Warning on Fake Prize Win Scams*, 2017).

An incident happened in March 2022, Khaleej Times reported that Abu Dhabi has launched a big-ticket draw in which there was a real winner. They tried to contact her, but she refused to claim her prize because she thought those were scammers that would steal her money (Kumar, 2022). After this incident, Big-Ticket has established a rule where they will publish the winner's name on their website and their social media account to make it easier for people to believe them.

3.1.9 Forensics Perspective:

There are steps that the users could follow to be able to distinguish between the fakeness of websites, pop-ups, and sweepstakes. If the user has a popup notification on his screen while browsing a regular website, then, before proceeding with any step, the user must follow some instructions that are advised by a forensic investigator (A. R. Ikuesan et al., 2020; R. A. Ikuesan, 2016; Zawali et al., 2021, Latto, 2021).

a. Check the website safety:

Use a website safety checker like Google Safe Browsing to quickly determine whether a website is trustworthy. This is a wonderful website safety-check tool since, according to Google, their website checker "examines billions of URLs per day looking for harmful websites". Another option is by using the VirusTotal online tool. It scans websites using over 80 antivirus engines and URL blacklisting services.

b. Browsers safety tool:

The most used web browsers available today have security features. These built-in browser capabilities can disable Flash content, prevent dangerous downloads, block intrusive pop-up adverts, and much more. To implement the above features, the steps of the respective browser include:

i. In Chrome: Settings > Advanced > Privacy & Security
ii. In Edge: Settings > Advanced settings
iii. In Firefox: Options > Privacy & Security
iv. In Safari: Preferences > Privacy

c. Check the URLs:

A common technique used to evaluate website safety is to check a URL before clicking on it. Before clicking on a link, find out where it leads, to determine whether it is secure.

Figure 5: Hovering a Mouse over a hyperlink.

In Figure 5, an investigator can determine fakeness by hovering the mouse over any link, the URL must show in the bottom left of the browser. On the web, most users simply scan text. Because of this, hackers frequently use visually similar characters to mislead you into visiting their phishing websites and unknowingly providing them confidential information (for example, "Yah00.com" instead of "Yahoo.com").

d. Checking the website HTTPS:

Another technique to ensure a website is secure is to verify that it uses HTTPS whenever you visit it. The primary protocol for transmitting data between your web browser and the websites you visit is HTTP (Hypertext Transfer Protocol). The "S" stands for "secure," and HTTPS is simply a secure version. Check for the padlock in the menu bar of your browser to discover if a site is HTTPS-encrypted. If you see it, the website you're on is encrypting your connection to it with a trusted SSL cryptographic certificate. Furthermore, combining HTTPS encryption with a virtual private network (VPN) will increase your security because the VPN encrypts all of your internet traffic from the time it leaves your device until it reaches the website you're viewing and back.

e. Don’t trust the “trust” badges:

Trust seals or logos typically appear on retail or e-commerce websites to indicate reliability. These icons stand out as evidence of legitimacy to customers who scan the website. Although many trustworthy websites utilize trust badges as well, they are unofficially recognized. These icons are frequently copied and pasted on websites without any meaningful security. A trust badge offers no information regarding the security procedures of the site in question.

4. Discussion

The mentioned hoaxes indicators in the study; clear signals, user manipulation, and fact-checking are three significant techniques to directly recognize a hoax. These three approaches are generic ways that any individual can implement when dealing with a stated fact. However, limitations would exist if it only depended on these approaches. Therefore, in the digital forensic field, two generic approaches were mentioned, human intervention and algorithm observation; the latter approach is an alternative. It analyzes data by using algorithms offer the ability to process and handle a big amount of data and avoid any kind of bias. For future work, this strategy can be improved by creating a digital tool that performs the same function but is only used by experts checking to prove the credibility of a statement.

Apart from this, the fakeness of websites due to cybersquatting is a hurdle for both technical savvies and general users. Throughout the investigation implemented in this subject, few facts were discovered. Other than the fact that the domain used for cybersquatting might be similar to the targeted individual or business domain, there’s a specific type that is incredibly malicious. Homograph attacks in cybersquatting not only make the domain similar to the legitimate domain but also makes it the same. The secret behind homograph attacks is that cybercriminals change conventional domain names, which typically consist of ASCII digits, letters, and special characters, into domain names that graphically resemble valid ones using Punycode, which is a subset of Unicode characters. Subsequently, these malicious fake websites would be extremely hard to detect and differentiate, perhaps even when using the known indicators mentioned in the investigation. To exemplify, a study has been attempted by a man named Xudong Zheng. Zheng has bought a domain named “xn--80ak6aa92e.com”, and if a user clicked on this link, they will be redirected to a fake website that looks exactly like Apple.com besides the fact that the domain in the URL would appear as Apple.com.

Moreover, depending on the expansion of sweepstakes, people now are facing a horrific time to distinguish between fake and legitimate sweepstakes. To figure out the types of sweepstakes, we have discussed earlier the indicators that a user must focus on to identify a sweepstake. If a fake sweepstake was determined by the indicators, several tools would help the user by checking the reliability of each source. The online tools will be a great opportunity to figure it out. However, some tools can't determine the fake sweepstakes. Therefore, still, people are facing fake sweepstakes daily since there is no proven tool to ban those sweepstakes from showing up. To add up, there is one practice a user can do that wasn't mentioned earlier, which is to block the pop-ups from showing in the browser. Indeed, it's not a long-term solution, however, could save some money for some users.

5. Conclusion

In conclusion, although fraudulent websites are a widespread threat to Internet users, many still have trouble recognizing them. Any unreliable internet site that is used to trick people into engaging in fraud or malicious attacks is a fake website. Perpetrators take advantage of the internet's anonymity to hide their genuine individual motives in numerous ways. While there are many valuable things an entity can do online, not everything is as it seems. Websites created for a variety of criminal objectives can be found among the millions of legal websites competing for users' attention. From misleading information to committing identity theft, these fake websites attempt it all. The observation presented in this study can be applied to diverse other genres of fake websites. However, further studies are required to develop a comprehensive indicator of fakeness. This is particularly important when novel genres are developed by malicious stakeholders. Furthermore, a formal benchmarking of the identification mechanism with industry standards is a major component of the future work of this study. Such a mechanism can be deployed as an API for websites as well as plugins agnostic to web browsers.

References

Abbasi, A., & Chen, H. (2009). A comparison of fraud cues and classification methods for fake escrow website detection. Information Technology and Management, 10(2-3 SPEC. ISS.), 83–101. https://doi.org/10.1007/s10799-009-0059-0

Abbasi, A., & Chen, H. (2013). *Detecting Fake Escrow Websites using Rich Fraud Cues and Kernel Based Methods*.

Agarwala, M., & Kang, S. (2021). Cybersquatting India: Genesis & Legal Scenario. *IJLMH*, *4*(5), 740–757.

Agten, P., Joosen, W., Piessens, F., & Nikiforakis, N. (2015). Seven Months' Worth of Mistakes: A Longitudinal Study of Typosquatting Abuse. *Proceedings 2015 Network and Distributed System Security Symposium*. https://doi.org/10.14722/ndss.2015.23058

Azorult Malware Information. (2019, December 20). Trendmicro. https://success.trendmicro.com/dcx/s/solution/000146108-azorult-malware-information?language=en_US

Baror, S. O., Ikuesan, R. A., & Venter, H. S. (2020). A defined digital forensic criteria for cybercrime reporting. Proceedings of the 15th International Conference on Cyber Warfare and Security, ICCWS 2020, ii, 617–626. https://doi.org/10.34190/ICCWS.20.056

Chan, J. (2022). Online astroturfing: A problem beyond disinformation. Philosophy and Social Criticism, 0(0), 1–22. https://doi.org/10.1177/01914537221108467Chandra, R., & Bhatnagar, V. (2019). Cyber-squatting: a cyber crime more than an unethical act. *International Journal of Social Computing and Cyber-Physical Systems*, *2*(2), 146. https://doi.org/10.1504/IJSCCPS.2019.100197

Chua, C. E. H., & Wareham, J. (2004). Fighting Internet auction fraud: an assessment and proposal. *Computer*, *37*(10), 31–37. https://doi.org/10.1109/MC.2004.165

Cybersquatting Examples: Everything You Need to Know. (2022, September 2). Upcounsel. https://www.upcounsel.com/cybersquatting-examples

Dinev, T. (2006). Why spoofing is serious internet fraud. *Communications of the ACM*, *49*(10), 76–82. https://doi.org/10.1145/1164394.1164398

Domain names and cybersquatting . (2017). *European IPR Helpdeask*.

Frischknecht, P. (2021). Detection of Cybersquatted Domains (Issue July). University of Bern, Switzerland.

Gyongyi, Z., & Garcia-Molina, H. (2005). Spam: it's not just for inboxes anymore. *Computer*, *38*(10), 28–34. https://doi.org/10.1109/MC.2005.352

Hewling, M. O. (2013). Digital forensics: an integrated approach for the investigation of cyber/computer related crimes. *University of Bedfordshire*.

Ikuesan, A. R., Salleh, M., Venter, H. S., Razak, S. A., & Furnell, S. M. (2020). A heuristics for http traffic identification in measuring user dissimilarity. Human-Intelligent Systems Integration, 2, 17–28.

Ikuesan, R. A. (2016). Online Psychographic Model for Insider Identification [Universiti Teknologi Malaysia]. http://portal.utm.my/client/en_AU/main/search/detailnonmodal?qu=Online+data+processing&d=ent%3A%2F%2FS

D_ILS%2F0%2FSD_ILS%3A851360~ILS~0&ps=300Karanicolas, M. (2020). The New Cybersquatters: The Evolution of Trademark Enforcement in the Domain Name Space. *SSRN*, *30*.

Kebande, V. R., Karie, N. M., & Ikuesan, R. A. (2020). Real-time monitoring as a supplementary security component of vigilantism in modern network environments. International Journal of Information Technology (Singapore). https://doi.org/10.1007/s41870-020-00585-8

Kumar, A. (2022, March 9). *UAE: Dh100,000 Big Ticket winner refuses to claim prize, thinks organiser's calls are fake* . Khaleej Times. https://www.khaleejtimes.com/uae/uae-dh100000-big-ticket-winner-refuses-to-claim-prize-thinks-organisers-calls-are-fake

Latto, N. (2021, September 24). *Website Check — How to Check Website Safety | AVG | AVG*. AVG. https://www.avg.com/en/signal/website-safety

Li, L., & Helenius, M. (2007). Usability evaluation of anti-phishing toolbars. *Journal in Computer Virology*, *3*(2), 163–184. https://doi.org/10.1007/s11416-007-0050-4

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research*, *15*(4), 336–355. https://doi.org/10.1287/isre.1040.0032

Mehta, I. (2020, March 11). *Hackers are using coronavirus maps to infect your computer*. The Next Web. https://thenextweb.com/news/hackers-are-using-coronavirus-maps-to-infect-your-computer

Mercer, J. D. (2000). Cybersquatting: Blackmail on the Information Superhighway. Boston University Journal of Science and Technology Law, 6(11).

Nangia, R. (2021). Cybersquatting From a Legal Perspective. *Legal Desire International Journal* , *8*.

Neil Chou, Robert Ledesma, Yuka Teraguchi, Dan Boneh, & John C. Mitchell. (2004). Client-side defense against web-based identity theft. *Computer Science Department, Stanford University, Stanford*.

Ntoulas, A., Najork, M., Manasse, M., & Fetterly, D. (2006). Detecting spam web pages through content analysis. *Proceedings of the 15th International Conference on World Wide Web*, 83–92. https://doi.org/10.1145/1135777.1135794

Orduña-Malea, E. (2021). Dot-science top level domain: Academic websites or dumpsites? *Scientometrics*, *126*(4), 3565–3591. https://doi.org/10.1007/s11192-020-03832-8

Petrosyan, A. (2022, July 20). *Leading industries with the largest share of domain name case filings due to cybersquatting as of June 2020*. Statista.

Police warning on fake prize win scams. (2017, November 12). Gulf News. https://gulfnews.com/uae/crime/police-warning-on-fake-prize-win-scams-1.2123257

Radhika Vivek Bhusari, & Karan Ramchandra Rampure. (2022). Cybersquatting: A Threat To The Globalising World. *Indian Journal of Law and Legal Research*.

Singh, H. P. (2018). Cyber Squatting and the Role of Indian Courts: A Review. *Amity Journal of Computational Sciences* , *2*(2).

Sousa-Silva, R. (2022). Fighting the Fake: A Forensic Linguistic Analysis to Fake News Detection. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, *35*(6), 2409–2433. https://doi.org/10.1007/s11196-022-09901-w

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it Hoax: Automated fake news detection in social networks. CEUR Workshop Proceedings, 1960, 1–12.

Tangermann, V. (2020). *Hackers Are Using Coronavirus Maps to Spread Malware*. Futurism. https://futurism.com/the-byte/hackers-coronavirus-maps-spread-malware

Yue Zhang, Serge Egelman, Lorrie Cranor, & Jason Hong. (2007). Phinding Phish: Evaluating Anti-Phishing Tools. *Carnegie Mellon University*.

Zahedi, F. M., Abbasi, A., & Chen, Y. (2015). Fake-website detection tools: Identifying elements that promote individuals' use and enhance their performance. Journal of the Association for Information Systems, 16(6), 448–484. https://doi.org/10.17705/1jais.00399

Zawali, B., Ikuesan, R. A., Kebande, V. R., Furnell, S., & A-Dhaqm, A. (2021). Realising a Push Button Modality for Video-Based Forensics. Infrastructures, 6(4), 54. https://doi.org/10.3390/infrastructures6040054

An Educational Scenario for Teaching Cyber Security Using low-cost Equipment and Open Source Software

Antonios Andreatos
Hellenic Air Force Academy, Division of Computer Engineering and Information Science, Dekeleia Air Force Base, Dekeleia, Attica, Greece

antonios.andreatos@hafa.haf.gr, aandreatos@gmail.com

Abstract: This work presents a set of hands-on educational activities designed to teach some cyber security concepts in the classroom. The experimental configuration used an ad-hoc wireless and wired network, and a Raspberry Pi implementing a Web and an SSH server. Students were connected using their own devices (laptops or tablets). Initially the students tested DoS attacks to the Web server using various tools. Next, the students had to create SSH accounts to the server and a pair of RSA keys; using their SSH accounts, the students had to transfer their public keys to the server. Finally, students had to attack the SSH service from Kali Linux running on virtual machines in teams, each team using a different tool. The scenario was implemented in parts during a series of lessons and was positively accepted and evaluated by the students, who got familiar with a number of concepts and tools of computer networking and network security. In the end, the students informally assessed the Kali Linux SSH attack tools. Ways to assess the students qualitatively and quantitatively based on their participation are also presented. These lab exercises used a series of open source software, as well as low-cost equipment.

Keywords: SSH Server, Web server, DoS attacks, SSH attacks, Kali Linux, Virtual Machines, Raspberry Pi, Network Security, Cyber Security Education.

1. Introduction - Computer networks and network security courses at the Hellenic Air Force Academy

The Hellenic Air Force Academy (HAFA) offers an undergraduate programme on Telecommunications and Electronics Engineering (TEE). This specialisation is equivalent to a Bachelor's degree in Electrical Engineering. At the fourth year of studies, the Division of Computer Engineering and Information Science offers the following courses to TEE students:

- Computer Networks I is offered during the 7th semester.
- Computer Networks II is offered during the 8th (last) semester
- Network Security (in essence, Cyber Security) is offered during the 8th semester.

Cyber Security requires a strong and wide background of computer engineering and information science courses ranging from computer networks, web technologies, database management systems, programming languages, software engineering, cryptography, operating systems and computer architecture, but also mathematics, probability and statistics, psychology, etc. (Andreatos, 2017). All of these courses are offered by our institution. However, academic Cyber Security textbooks emphasise on theory and underlying principles, rarely focusing on particular systems and practical issues (Yurcik and Doss, 2000). For this reason, several lab exercises complete the theoretical education.

During the 3rd year of studies a background course of Internet technologies using the LAMP stack is taught. (LAMP stack is a bundle of four different software technologies used to build websites and web applications. LAMP is an acronym for the Linux operating system, the Apache web server, the MySQL database management system and the PHP programming language.) Procedural and object-oriented programming courses are offered in the 1st and 2nd years respectively. Operating Systems are taught at the 3rd year of studies. All these and other courses offered by other divisions are prerequisite courses for Cyber Security.

In our Network Security course the material is offered in lectures, class demonstrations, assignments, lab exercises, case studies and other activities such as videos and movies (**Andreatos, 2020).** Junior and senior students also participate in annual national and international cyber defence exercises.

2. Related works

Teaching cyber security is very different from teaching regular courses; therefore, a wide variety of teaching methodologies have been implemented during the past 15 years. Some common methodologies are: demonstrations (capture of network traffic, network simulation, video clips, movies), hands-on techniques including exercises in real and virtual labs (Prvan and Ožegović, 2020; Surma, 2003), serious games (Hwang and

Helser, 2022), and case studies (Cai, 2018). Two of the most recent surveys on the subject are: the research on Teaching Computer Networks by Prvan and Ožegović (2020) which included 196 references from ACM Digital Library and IEEE Xplore Digital Library; and the research on Teaching Cyber Security Course in the Classrooms by Churi and Rao (2021) which included 35 quality papers on innovative pedagogy on Cyber Security and similar courses from IEEE, Springer, Elsevier, and some other quality journals.

Churi and Rao (2021) support that Cyber Security is a practical subject which requires:

- Hands-on simulations of attacks and countermeasures;
- Case studies on current trends of Cyber Security;
- Use of open source software (OSS) and tools.

Yu (2007) presented a security lab setup, rules to avoid potential illegal activities for security considerations, the grading policy and the description of a term project including several successive exercises. The need for hands on laboratory exercises is emphasised by Erickson and Kim (2021), which support that "the ethical challenges of teaching students how to hack systems is real, but so is the need for them to see what is happening from both sides of the table to be able to better defend the attacks when they are in the role". One way to achieve this is to teach hacking techniques to students. Teaching students attack methods and tools in a controlled lab environment, enables them to learn what hackers are targeting and therefore find ways to defend their organisation against such attacks more easily than if they were just working on cyber defence techniques (Erickson and Kim, 2021).

Reports are indispensable in this type of lab work. The students should provide a report on what attacks were attempted, how likely they believe a skilled administrator would have detected their attack, and how the attack could have been prevented (Yu, 2007).

3. The educational scenario

3.1 Educational objectives

During the 2021-22 academic year, in an effort to promote hands-on Cyber Security education, the teacher of Computer Networks and Network Security designed a 10-12 class hours project including several successive hands-on exercises (hereafter called "scenario"). The scenario used OSS, low-cost equipment and student devices. The educational objectives of the scenario were:

a) To teach the following Computer Networking principles in practice:

- How to set up an ad-hoc LAN;
- How to monitor LAN users and network traffic;
- How to connect to a remote machine using SSH. ‘SSH’ stands for "Secure Shell", a protocol designed as a secure alternative to unsecured remote shell protocols (SSH, 2023);
- How to use virtualisation; how to connect a virtual machine (VM) to a LAN.

b) To teach the following Network Security principles in practice:

- How to generate a pair of RSA keys for public-key encryption;
- How to copy a user’s public key to a server and disable username/password authentication;
- What are the configuration files of the SSH server (sshd_config, known_hosts) and how to access them;
- How to launch a DoS attack against a web server (Williams, 2021);
- How to prepare a dictionary for a brute-force attack;
- How to brute-force attack an SSH server using a dictionary.

Step-by-step instructions, bibliography and the objectives of the lab exercises were communicated to the students in advance in two assignments (in November 2021 for Computer Networking I and in March 2022 for Network Security). The students were also advised to take notes and screenshots during the lab to put them in their reports. Figure 1 presents the steps of the scenario.

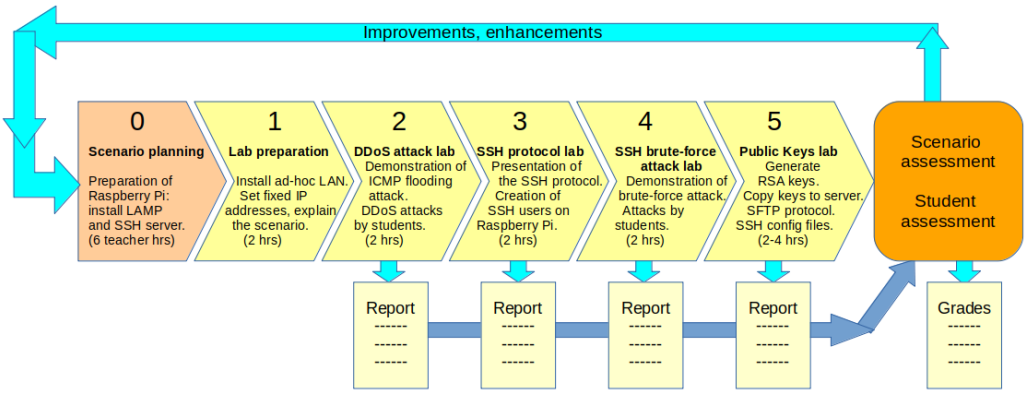


Figure 1: The educational scenario step-by-step

The whole activity was implemented during lecture hours, approximately two hours per week, in parallel with Computer Networking and Network Security theoretical classes, so that the students had enough time to get familiar with the subject. Several freely available software tools were used in these exercises, including Angry IP Scanner (for Linux), Advanced IP Scanner (for Windows), ssh-keygen, PuTTY (for Windows), VirtualBox, etc. Wireshark open-source packet analyser was used to capture network traffic in order to monitor students' activity (Ghafarian, 2014). Patator, Medusa and Metasploit were the Kali Linux tools used to attack the SSH server.

During the lab exercises the instructor provided guidance and technical support; he used tools and ways to objectively assess students based on their participation, which are also presented in this paper. The students assessed the scenario positively; in addition, it was found that, in contrast to theoretical lectures, such alternative educational activities support a variety of learning styles.

Finally, it should be noted that the planning of the scenario, the preparation of each lab, as well as the students' assessment, required many hours of work on the part of the teacher.

This paper describes the educational scenario in details in order to facilitate colleagues who may want to try it in their classroom.

3.2 Scenario planning

Due to its long duration, the activity was implemented in parts and spanned several weeks from both the Fall 2020 and Spring 2021 semesters. For this reason, the equipment had to be minimal, portable and fast to set up.

In the Fall 2020 semester the scenario was executed during the **Computer Networks** I hours; in Spring 2021 it was executed during the **Computer Networks** II and Network Security hours, in parallel with theoretical lectures, so that the students had enough time to get familiar with the prerequisite theoretical background. The three above courses are taught by the same teacher, so he had the flexibility to set the background, avoid overlays and gaps, and synchronise the lab exercises between classes. Table 1 summarises the scenario activities and the tools required in each step.

Table 1: Scenario activities and tools

No.	Activity	Tools
1	Create an account on Raspberry Pi	Terminal & ssh command
2	Perform a DoS attack on Raspberry Pi	LOIC and HOIC, hping3, Python scripts
3	Check site availability	Browser
4	Detect and present the specific attacks	Wireshark, tcpdump, netstat
5	SSH Brute Force attack on Raspberry Pi	Patator, THC Hydra, Medusa, Metasploit
6	Detect, sniff, record and present the above attack	Wireshark, tcpdump, netstat, nmap
7	Create a pair of RSA keys	ssh-keygen
8	Upload the public key to the server	ssh-copy-id, sftp, PuTTY
9	Configure the ssh-server to require keys for connection, and prevent root login	sshd_config

Activities 1 to 6 were implemented in December 2021. Students had to submit a report with their activities. Activities 7 to 9 were implemented in March 2022 (in the 2nd semester, Network Security course) after teaching the section about asymmetric encryption; students had to submit another report. Table 2 presents most tools used in the proposed scenario organised in categories.

Table 2: A taxonomy of tools used according to their function

Monitoring tools	Connection tools	DDoS tools	SSH attack tools	Discovery/ Network traffic	Support tools
Angry IP Scanner	ssh	ddos.py	THC Hydra	nmap	VirtualBox with extension pack
Advanced IP Scanner	PuTTY	LOIC and HOIC	Patator, Medusa, Metasploit	Wireshark	kali Linux image for VirtualBox
Bash commands	sftp	hping3	Dictionary tools	tcpdump	ssh-keygen

4. Lab execution

4.1 Equipment and infrastructure

Raspberry Pi is a tiny, low-cost but powerful computer board, based on a 'system on a chip'; it was first released in February 2012 (https://www.raspberrypi.org). Figure 2 shows a Raspberry Pi 4 Model B (2019). To date the Raspberry Pi Foundation has produced several generations of boards with several models which cover a wide range of open-source projects. Raspberry Pi can run several operating systems which are offered as ready-made images for downloading, with preinstalled software which converts Pi to a Print Server, Media Server, Game Sever, LAMP Web Server with WordPress, etc. In this work the Raspbian operating system has been used. Raspbian is a free version of Debian optimised for the Raspberry Pi hardware (https://www.raspbian.org). Raspberry Pi's use an SD card in order to store the operating system, applications and additional software installed.

Figure 2: Raspberry Pi model 4

A Raspberry Pi model 3B was used in these lab exercises. An ad-hoc wireless Local Area Network (LAN) was set-up to support the whole scenario. The Raspberry Pi was used as a 'headless' server (i.e., without keyboard and monitor); its software was prepared by the teacher. The web server was Apache and it hosted a site from 3rd year's course "Internet Technologies", familiar to the students. Because this device does not avail a Wi-Fi port, it was connected via wire (Ethernet port) to a Wi-Fi range extender used in access point mode. Students' and instructor's laptops were connected via Wi-Fi to the ad-hoc LAN. The Wi-Fi Range extender device was a TP-Link AC750. It operates at 2.4 GHz and offers a data exchange rate of 300 Mbps.

The recommended operating system for students was Kali Linux, either installed on bare hardware or in a virtual machine. Figure 3 presents a photograph from the lab setup.

Figure 3: A photograph from the lab showing the Raspberry Pi and the projector

4.2 Instructor's role

The whole experiment consumed a lot of the teacher's time and effort in the design, implementation, student support and assessment. The instructor was guiding the students through the scenario, demonstrating the steps, offering technical support, as well as displaying the big picture through a projector (for instance, the results of the students' activities on the Raspberry Pi or on the network traffic).

The instructor had also to keep track of students' activities in order to make sure that no one was left behind, so that all students would be able to finish the activity. Special attention was paid to synchronising the activities with the background and skills of the students.

4.3 Procedure

The students were instructed to install the required software which included:

- Network monitoring tools (Advanced IP Scanner, Angry IP Scanner, nmap)
- Network traffic / packet capture tools (
- DoS attack tools (LOIC, HOIC and hping3)
- Virtualisation tools and a Kali Linux image (VirtualBox)
- SSH brute force attack tools (embedded in Kali).

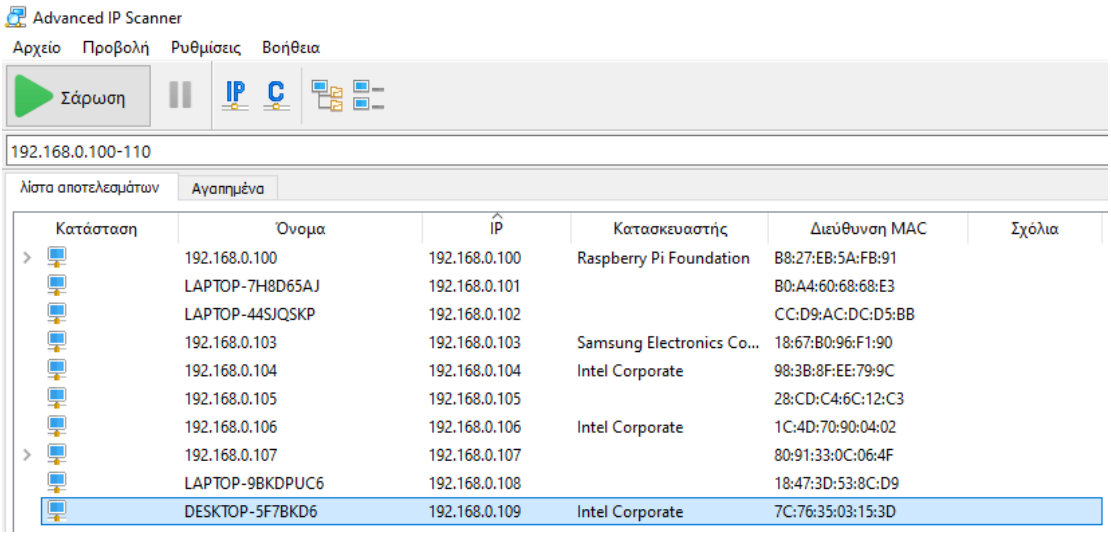


Figure 4: A screenshot from Advanced IP Scanner

The teacher used a Linux laptop to connect to the Raspberry Pi via SSH. From this laptop the teacher was also monitoring and recording students' activities using proper software (Angry IP scanner, Advanced IP Scanner, Wireshark and tcpdump). Figure 4 presents a screenshot from Advanced IP scanner.

DoS/DDoS attacks

The teacher launched an ICMP flooding attack with endless ping and displayed the traffic on screen using tcpdump (Figure 5). Afterwards he displayed the students' traffic from Wireshark on the projector.

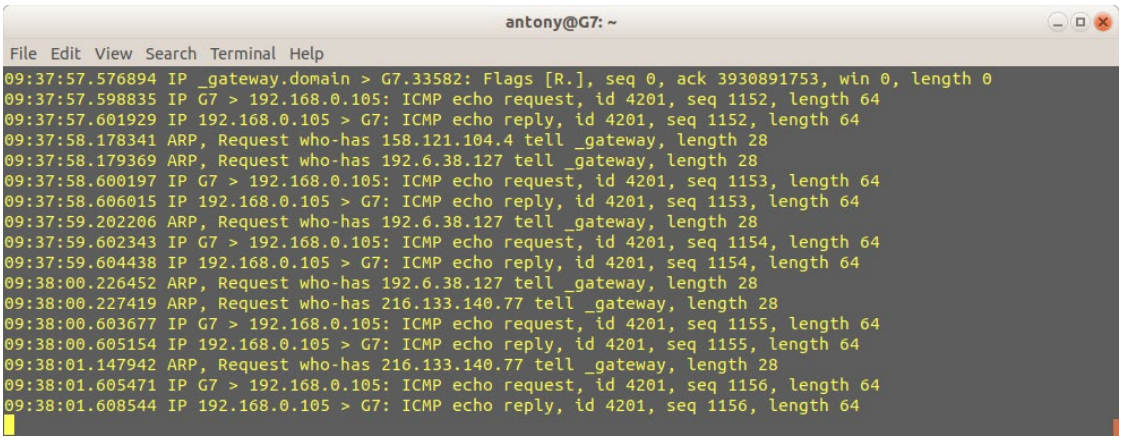


Figure 5: Screenshot from an ICMP flooding attack demonstrated by the teacher

LOIC and HOIC tools were proposed for DoS attacks on the Raspberry Pi's web server. However, a student had some Python script tools which shared with his classmates. Next, the students started the DDoS attack from their devices. Figure 6 demonstrates a screenshot from a student's report demonstrating a DoS attack via 'ddos.py'. The message on the left writes that 'access to this site is not possible' in Greek.

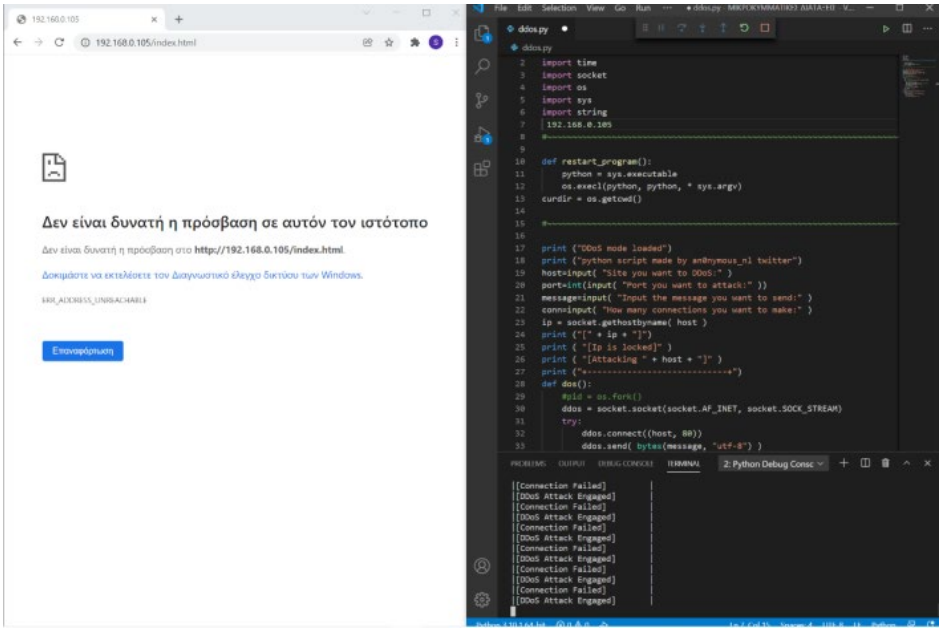


Figure 6: Screenshot from a student's report demonstrating DoS attack via 'ddos.py'

Creation of SSH users

The teacher demonstrated how to connect to Raspberry Pi as root and how to create a new SSH user with username and password. Then, the students created personal SSH accounts using usernames related to their lastname. Next, the students connected to Raspberry Pi using the username and password they created and explored their home directory.

The teacher checked their connection via the 'w' command. The 'w' command is a Linux terminal built-in tool that allows administrators to view information about currently logged in users. This includes their username, where they are logged in from (IP address in our case), log-in time, what they are currently doing, etc. This is a milestone used in quantitative assessment. Figure 7 demonstrates a representative screenshot.

```
andreatos60@raspberrypi:~ $ w
 11:52:11 up 35 min, 12 users,  load average: 0.10, 0.13, 0.08
USER     TTY      FROM             LOGIN@   IDLE   JCPU   PCPU WHAT
pi       tty7     :0               11:17   34:53   3.48s  0.55s
/usr/bin/lxsession -s LXDE-
pi       tty1     -                11:17   34:53   0.56s  0.42s -bash
[redacted] pts/1    192.168.0.107    11:39    0.00s  0.42s  0.04s w
[redacted] pts/2    192.168.0.111    11:47    1:19   0.38s  0.38s -bash
[redacted] pts/3    192.168.0.109    11:45    6:10   0.37s  0.37s -bash
[redacted] pts/4    192.168.0.101    11:46    2:19   0.42s  0.42s -bash
[redacted] pts/5    192.168.0.104    11:46    1:18   0.45s  0.45s -bash
[redacted] pts/6    192.168.0.110    11:46   51.00s  0.36s  0.36s -bash
[redacted] pts/7    192.168.0.113    11:47    1:01   0.45s  0.45s -bash
[redacted] pts/8    192.168.0.114    11:48    1:15   0.46s  0.46s -bash
[redacted] pts/9    192.168.0.103    11:51    1:06   0.36s  0.36s -bash
[redacted] pts/10   192.168.0.106    11:51   35.00s  0.43s  0.43s -bash
```

Figure 7: A screenshot from 'w' command demonstrating students connected to Raspberry Pi

The teacher recorded the network traffic with Wireshark (Figure 8). From the Wireshark Statistics menu the teacher could identify their activity (based on the student's MAC and IP addresses). This was another milestone for quantitative assessment.

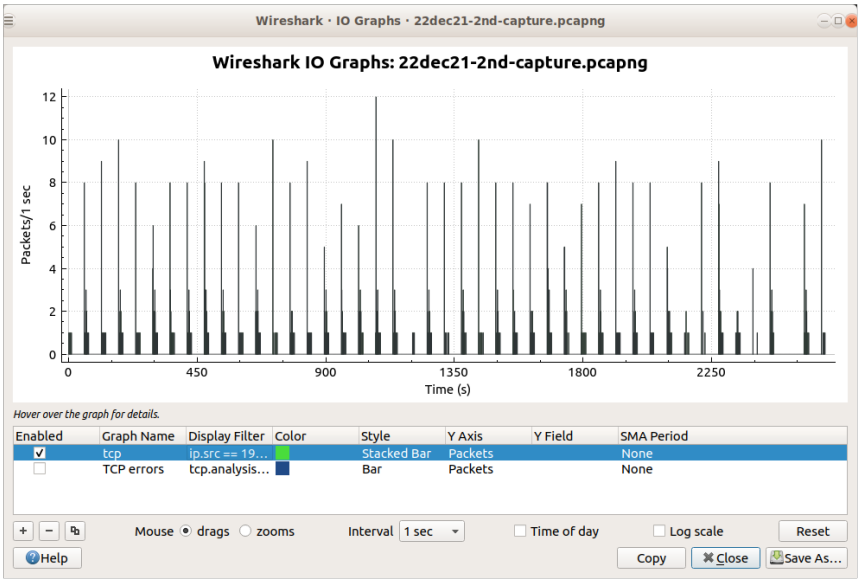


Figure 8: Screenshot from Wireshark demonstrating the SSH brute force attack

Brute force attack on the SSH server

The SSH brute force attack (CMU, 2023) took place in December 2021. Initially the teacher demonstrated an attack using Hydra and then he asked the students to do the same using their own tool (Patator, Medusa, Metasploit) in teams. Three teams were formed from the eleven students. Each team had to find instructions on how to use their tool. Finally, each team should assess the usability of their tool. SSH brute force attack tools usually include a default dictionary with common usernames and passwords. The teams had to add their classmates' usernames and possible passwords related to their personalities (i.e., use social engineering). Finally, the SSH brute force attacks succeeded. Figure 9 presents a screenshot from the SSH brute force attack using Metasploit.

4.4 Evaluation of the SSH brute force tools

The teams, after using their assigned tool, commented on its efficiency and usability. Medusa and Metasploit got a satisfactory assessment; Patator disappointed their users.

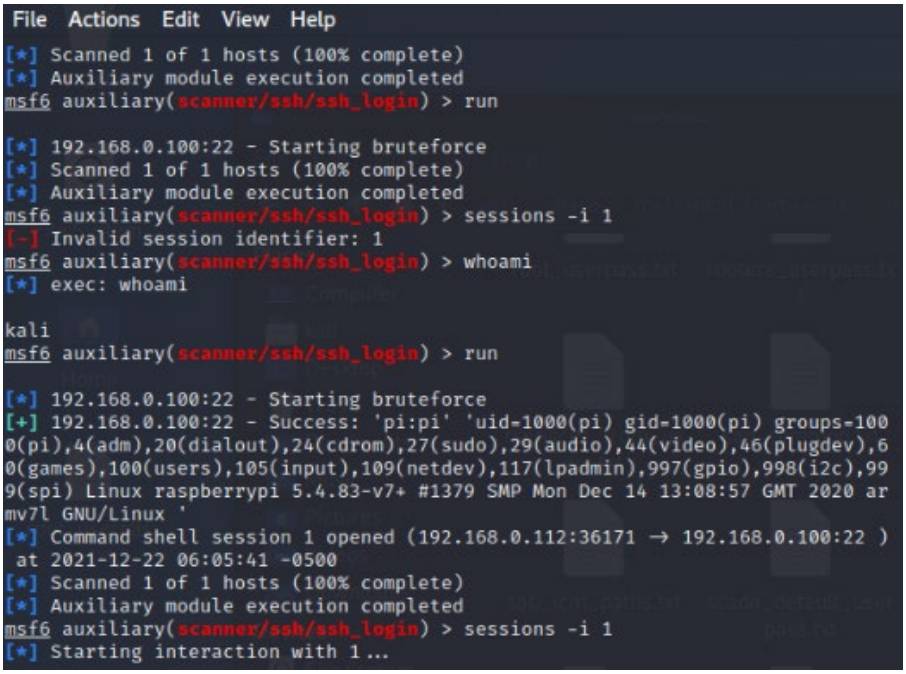


Figure 9: Screenshot from the SSH brute force attack using Metasploit

4.5 Use of public keys

After this date followed the Christmas holidays and the final exams of the Fall semester. The scenario continued in the Spring 22 semester, during the Computer Networks II and Network Security courses.

Asymmetric cryptography was taught in Network Security and after that, the corresponding lab exercise (RSA key generation) was implemented in March 2022. The students had to transfer their public keys to the SSH server in order to enable automatic connection and disable username and password login in an effort to reduce the attack surface and make the SSH server immune to brute force attacks.

5. Discussion

5.1 Achieving the learning objectives

In this series of hands-on lab exercises students had the opportunity to interact with real networking hardware and use a variety of tools and methods. The students had to prepare their own devices (laptops and tablets) by installing software. They had the opportunity to try a lot of tools in practice and get valuable practical experience; they also had the chance to see the results of their activities in real time on screen. The scenario kept the students interested till the end; they cooperated harmoniously both with the teacher and among themselves, since some activities were implemented in teams.

From their reports we evaluated the students' achievements which are listed below:

- The students were able to set their IP address, as well as find the IP addresses of the Raspberry Pi and the other users.
- They were able to create a username και password for connecting to the Raspberry Pi.
- They performed a DoS attack on the Raspberry Pi's web server using several tools.
- They performed a brute force attack on the Raspberry Pi's SSH server in teams.
- They found instructions on how to use their tool and assessed the usability of their tool in the end.
- They created a pair of RSA keys and copied their public key to the Raspberry Pi's SSH server.
- Only one out of eleven students managed to verify that his public key was there, as well as, viewed the configuration files of the SSH server using the sftp protocol.

Students' comments were positive; an interview with the aforementioned student who prepared the best report revealed that he gained valuable practical experience and enjoyed the scenario much more than theory. Hence, we conclude the success and the effectiveness of this educational activity.

5.2 Evaluating the students

Ways to assess students both qualitatively and quantitatively, based on their participation and achievements, are presented here. We begin with the quantitative metrics. The following milestones were assessed:

1) Assessment of DoS attacks: each student had to use one of the suggested DoS attack tools against the Web server of Raspberry Pi, until a proper message showing that the page was not available appeared on their screen. The instructor could observe the packet floods in the packet capture (given the students' IP and MAC addresses).
2) The creation of SSH user accounts in the Raspberry Pi SSH server was verified by exploring the students' home directories.
3) The brute-force attacks on the SSH server were verified through the Wireshark packet capture files.
4) The students had to submit a report with their activities as an assignment in the Computer Networks course, describing their activities and supporting them with screenshots.
5) Generating an RSA key pair and uploading the public key to the server was verified by exploring the students' home directories in Raspberry Pi.
6) The students had to submit another report as an assignment in the Network Security course.

The teacher could check student activity during the various steps of the scenario in real time using a variety of tools and methods including monitoring tools (Angry IP Scanner or Advanced IP Scanner), packet captures and Linux terminal commands.

Qualitative assessment was based on the teacher's observations on student participation and interest, as well as their roles in the team activities (leaders, followers or observers).

5.3 Future work

In the future we intend to improve and extend the scenario including additional activities such as:

1) The use of Wireshark, tcpdump and netstat by the students for recording and documenting their attacks.
2) The use of Firewall (iptables, ufw) on Raspberry Pi.
3) The use of Fail2Ban and Snort (**Ghafarian, 2014**) on Raspberry Pi.

We also plan to add more questions to the assignment.

6. Conclusion

This paper presented a cyber security scenario implemented in class using OSS, low-cost infrastructure and students' devices. The students saw how to create an ad hoc Wi-Fi LAN, connect a Raspberry Pi, launch DoS attacks using various tools, monitor the user activities using various tools, capture, monitor and analyse the network traffic, etc. The students got familiar with the Kali Linux platform and virtualisation, created a pair of keys and uploaded the public key to the SSH server so that they could login without username and password; they used three brute force attack tools against the SSH server, as well as assessed these tools.

This educational scenario gave practical lab experience and inspiration to the students, one of which completed the Massive Open Online Course (MOOC) "Networking Essentials" after graduation. Students with practice-oriented learning styles had the opportunity to learn in their own way. So this scenario was used as an alternative assessment activity, instead of exams and tests (**Andreatos and Leros, 2023**). The advantages of the scenario presented are:

- It uses low-cost, off-the-shelf equipment.
- It uses OSS and a variety of tools.
- The experimental infrastructure is inexpensive, portable and fast to install.

The current study verified that to teach technologically advanced topics such as cyber security, innovative teaching and assessment methods need to be applied (**Churi and Rao, 2021).** In the future we plan to improve the scenario in order to include more activities and tools.

Acknowledgements

The author wishes to thank the reviewers for their constructive feedback.

References

Andreatos, A.S. (2017) “Designing educational scenarios to teach network security”, IEEE Global Engineering Education Conference (EDUCON), pp 1606-1610. https://ieeexplore.ieee.org/document/7943063

Andreatos, A. (2020) “Movies as an Aid to Teach Principles of Cybersecurity and Cybercrime in Higher Education”, International Journal of Education and Information Technologies, Vol 14, 2020, pp 76-82.
http://www.naun.org/main/NAUN/educationinformation/2020/a202008-010(2020).pdf, http://doi.org/10.46300/9109.2020.14.10

Andreatos, A. and Leros, A. (2023) “Can Oral Grades Predict Final Examination Scores? Case Study in a Higher Education Military Academy”, accepted for publication in MDPI Analytics.

Cai, Y. (2018) "Using Case Studies To Teach Cybersecurity Courses", Journal of Cybersecurity Education, Research and Practice, Vol 2018, No. 2, Article 3. https://digitalcommons.kennesaw.edu/jcerp/vol2018/iss2/3

Churi, P. and Rao, N.T. (2021) “Teaching Cyber Security Course in the Classrooms of NMIMS University”, International Journal of Modern Education and Computer Science (IJMECS), Vol 13, No. 4, pp 1-15. https://www.mecs-press.net/ijmecs/ijmecs-v13-n4/IJMECS-V13-N4-1.pdf

Ghafarian, A. (2014) "An Empirical Study of network forensics analysis tools", Proceedings of the 9th International Conf. on Cyber Warfare and Security, West Lafayette, Indiana, USA, pp 366-370.

Erickson, M. and Kim, P. (2021) “Designing cybersecurity curriculum: Exploring the need for industry certifications and experiential learning”, Issues in Information Systems, Vol 22, Issue 4, pp 9-20. DOI: https://doi.org/10.48009/4_iis_2021_9-21

Hwang, M.I. and Helser, S. (2022) "Cybersecurity educational games: a theoretical framework", Information and Computer Security, Vol 30 No. 2, pp 225-242. https://doi.org/10.1108/ICS-10-2020-0173

CMU Information Security Office, Protect Against Brute-force/Dictionary SSH Attacks,
https://www.cmu.edu/iso/aware/be-aware/brute-force_ssh_attack.html

Prvan, M. and Ožegović, J. (2020) “Methods in Teaching Computer Networks”, ACM Transactions on Computing Education (TOCE) 20, pp 1–35. https://dl.acm.org/doi/10.1145/3394963

Surma, D. R. (2003) "Lab exercises and learning activities for courses in computer networks", 33rd Annual Frontiers in Education, FIE 2003, Westminster, CO, USA, 2003, pp T2C-21, doi: 10.1109/FIE.2003.1263297.

What is SSH? (2023) [Online]. Available at https://phoenixnap.com/kb/what-is-ssh [accessed 7 March 2023].

Williams, L. (2021) “10 Best FREE DDoS Attack Tool Online (2023)”, Updated January 7, 2023 [Online]. Available at https://www.guru99.com/ddos-attack-tools.html [accessed 7 March 2023].

Yurcik, W. and Doss, D. (2000) “Information Security Educational Initiatives to Protect E-Commerce and Critical National Infrastructures”, Information Systems Education Conference (ISECON), Philadelphia, PA, USA.

Yu, Y. (2007) “Designing hands-on lab exercises in the network security course”, J. Comput. Sci. Coll. 22, 5, pp 105–110.

A Methodical Framework for Conducting Reconnaissance and Enumeration in the Ethical Hacking Lifecycle

Fouz Barman, Nora Alkaabi, Hamda Almenhali, Mahra Alshedi and Richard Ikuesan
Computing and Applied Technology Department, College of Technological Innovation, Zayed University, Abu Dhabi, United Arab Emirates

fouz.barman@hotmail.com
Noraalkaabi2001@gmail.com
Hamdaalmenhali@hotmail.com
mahralshedi2019@gmail.com
richard.ikuesan@zu.ac.ae

Abstract: Reconnaissance and enumeration are both equally significant phases of the penetration testing lifecycle. In hindsight, both reconnaissance and enumeration seem to be very similar as the pair involve information gathering. Whilst reconnaissance leverages passive approaches without direct interaction with the target, enumeration exploits susceptibilities and vulnerabilities in direct client-server communication. Both phases involve gathering information and pinpointing the attack surface within the network of the target. To do so, powerful tools such as Nmap and Netcat are utilized by ethical hackers and penetration testers to identify and resolve security vulnerabilities and weaknesses. Nmap is an open-source command-line tool used for information gathering, network discovery, and security auditing. Whereas Netcat is a back-end tool that manages networks, monitors traffic flow between systems, as well as allows port scanning and listening. However, the plethora of tools and approaches available for these two phases often introduce inconsistencies and time wastage, which can lead to frustration and poor outcome for inexperienced penetration testers. Additionally, not all commands found online are relevant and applicable. In such situations, there is a high probability that the user will feel overwhelmed and exasperated with the overflow of new and foreign information. To address this daunting challenge, this study developed a methodical framework that can provide a technical guide for the reconnaissance and enumeration phases of the penetration testing lifecycle. Furthermore, a clear and thorough step-by-step procedure and detailed explanations of each stage and commands initiated using Nmap and Netcat are provided. The output of this study will be extremely beneficial and informative to a vast group of audience, ranging from university students majoring in security to individuals interested in ethical hacking, and even someone looking for a job with a position of a penetration tester. Furthermore, this technical guide on Nmap and Netcat extends the common body of knowledge in penetration, as a bridge between the industry and academia.

Keywords: Penetration testing, Penetration testing framework, reconnaissance, enumeration, ethical hacking, Nmap, and Netcat.

1. Introduction

Ethical hacking, also known as white hat hacking, is the utilization of hacking techniques to identify vulnerabilities in computer systems, networks, and web applications to improve security. This proactive approach helps to identify and fix security weaknesses before they can be exploited by malicious actors. Furthermore, ethical hacking includes penetration testing, simulating an attack on a computer system, network, or web application to identify vulnerabilities and assess the effectiveness of security measures. Ethical hackers use the same tools, techniques, and tactics as malicious hackers, however, with the permission of the system or network owner and the intention of making the system or network more secure (Kebande, Karie, and Ikuesan 2020). Historically, ethical hacking can be traced back to the 1970s when the US government began to employ hackers to test the security of their systems. This practice was formalized in the 1980s with the creation of the US Air Force's Computer Emergency Response Team (CERT). As the internet became more widely used in the 1990s, private companies also began to hire ethical hackers to test their systems. Today, many professional organizations and certifications specialize in ethical hacking, such as the offensive security certified professional, International Association of Computer Science and Information Technology (IACSIT) and the Certified Ethical Hacker (CEH) certification, SANS GIAC web application penetration testing, as well as other non-prominent organizations. A summary of organizations that provide certifications to ethical hacking, penetration testing as well as other forms of offensive proactive security is further highlighted in word-cloud depicted in Figure 1.

Figure 1: Information Security Certification Word Cloud

Cybersecurity can be defined in many ways, however, it is always defined as the organization and gathering of resources, processes, and structures used to defend cyberspace and cyberspace-enabled systems from occurrences are known as cybersecurity (Craigen, Diakun-Thibault, and Purse 2014). Controlling access to data on the systems became a major point of worry when these time-sharing systems arose in the mid to late 1960s and many more jobs were using the web. Furthermore, the cybersecurity scan started in the 1970s after the researcher Bob Thomas designed and produced a computer program named Creeper that can move in the ARPANET's network. Moreover, the inventor of the email Ray Tomlinson designed a program named Reaper that is designed to capture and delete Creepers. The Reaper is one of the first antivirus malware-scanning software that is self-replicating. In the year 1987, the commercial antivirus software where Kai Figgie Andreas Luning released their antivirus (GeeksforGeeks 2022). Moving on to the early 2000s, criminal groups began to heavily fund professional cyberattacks, while governments started to crack down on the criminality of hacking, handing out large sums of money to responsible hackers. Information security continues to progress as the internet increases, but so do viruses (GeeksforGeeks 2022). Cyberattacks have been increasing during the past years which ensures that organizations and individuals must include mechanisms and defense systems to reduce their risk of them. The defense mechanism used can be multi-factor authentication to avoid malicious actors accessing the user account. In addition, creating long complex passwords and changing them regularly to avoid password attacks. Moreover, malware scanners are low-cost tools to scan devices for vulnerabilities in the devices. Additionally, using a VPN-capable firewall which is hardware to encrypt all communication between the device to reduce the risk of brute force attacks can also work as a filter to block access to malicious websites. Moving on to cyberattacks, the topmost common attacks are DoS Denial of Service, Phishing, SQL injection, Ransomware, and XSS Cross-Site scripting. DoS attacks are attacks that overwhelm a target to stop it from functioning. Phishing attacks are attacks that visually look like they are from a trusted party, however, it includes malicious code in them. SQL injection attacks are attacks that target websites that use a database to service their clients. SQL injection is when an attacker injects malicious code into the user input to gain access or modify code. Ransomware attacks are attacks where the victim's system is held captive by ransomware until they agree to pay the attacker a ransom. After receiving payment, the attacker instructs the target on how to retake control of their computer. The target of a ransomware attack is to download ransomware from a website or an email attachment. The malware is designed to take advantage of flaws that have not been addressed by either the system's vendor or the IT staff (A. Singh, Ikuesan, and Venter 2019). After that, the ransomware encrypts the target's computer. Ransomware can often be used to target several parties by limiting access to many machines or a central server critical to corporate operations. Cross-Site Scripting attacks XSS is when the attacker creates malicious scripts and transmits them into clickable content that is sent to a target victim browser. Once the victim clicks on the content, the malicious script will run on the victim's device.

2. Background

Penetration testing and ethical hacking are essential in the information security lifecycle for several reasons. Firstly, it helps to identify vulnerabilities in systems and networks before they can be exploited by malicious hackers. By identifying and addressing these vulnerabilities, organizations can prevent cyber-attacks and protect

sensitive information, such as financial data and personal information. Secondly, ethical hacking is also important for compliance, as many industries are subject to regulations that require regular security testing. Thirdly, it also helps organizations to improve their overall security posture and stay ahead of the ever-evolving threat landscape. Furthermore, ethical hacking has been widely adopted by organizations as a key security strategy. In today's digital age, organizations are constantly at risk of cyber-attacks, and they must have the necessary security measures in place to protect their networks and data. Organizations that conduct regular ethical hacking and penetration testing are better prepared to detect and respond to cyber threats and are less likely to suffer data breaches and other security incidents. Fundamentally, the ethical hacking lifecycle contains five phases which are reconnaissance, enumeration, exploitation, post-exploitation, and clearing of tracks. The first two phases often play a critical success factor in any hacking exercise. However, this is often fraught with inconsistencies and subjective processes.

Given that some ethical hackers often lack formal training, experience, and understanding, ethical hacking has been susceptible to a diverse range of problems. The first problem relates to accessibility. Typically, ethical hackers need both physical and logical access to the target organization to conduct an assessment. In essence, the ability of ethical hackers to carry out their tried-and-true assaults depends on the access rights and privileges provided by the organization. This could be within the black, white, or grey box principle. The ability to effectively manage the first two phases of the hacking lifecycle is quintessential to the success of the process, especially with the least privileges (Yaacoub et al. 2021). Secondly, the problems regarding security and privacy must be taken thoughtfully, since organizations can be under simulated attacks, and their data, information, system privacy, integrity, and confidentiality are being targeted as well. Consequently, ethical hackers are severely required to explain, demonstrate, and document each phase performed. To provide a reliable baseline for developing trust and ensuring compliance, this study developed a conceptually practicable framework for the management of the first two phases of the ethical hacking cycle. The proposed framework is presented in the next section.

3. Proposed Framework

As a step towards enhancing the penetration testing process, this study presents a process model that can be used to reduce the daunting and confusing process of the first two stages in the penetration cycle. This is further illustrated in Figure 2. The process model details the process of managing the reconnaissance (reconn) and the enumeration (enum) phases in an agnostic alignment. The proposed management framework entails the process of ensuring ethical compliance during the reconnaissance and enumeration phases. Given that most ethical hacking process leverage diverse open-source tools and intelligence, the proposed framework further specify the potential platforms such as the use of Parrot OS and Kali Linux as the hacking platform. Within this platform, several off-the-shelf tools and open-source tools exist while others can be installed. To align with the principle of ethics, and the need for transparency in the process, the proposed framework integrates documentation and justification into every aspect of both phases. This includes the core process of curating the attack surface on the target based on the extensive use of the hacking platform. An implementation of this proposed framework is presented in the subsequent sections. Firstly, the methodology employed is discussed and then followed by the result obtained.

4. Methodology

This section of the report covers the methodology and steps undertaken to ensure a successful ethical hacking exercise. Additionally, this study highlights the two most important phases of the entire process, the reconnaissance and enumeration phases. To implement this process, kali-linux 2022.4-amd64 [1]was installed on a core i7 12GB RAM Windows 10 computing running using the virtual box as the virtualization technology. The default configuration for the Kali Linux platform with username and password as ‘kali’ was used. Furthermore, the victim machine was downloaded from Vulhub[2] open-source vulnerable machine repository. To demonstrate the documentation process, each step followed was carefully recorded using a combination of notes, and saved output (example, nmap –sP 192.168.170.0/24 >>output.txt). A synopsis of some common commands and their corresponding interpretation is further presented in Tables 1 and 2 for Nmap and Netcat respectively.

[1] https://www.kali.org/get-kali/#kali-virtual-machines
[2] https://www.vulnhub.com/entry/napping-101,752/

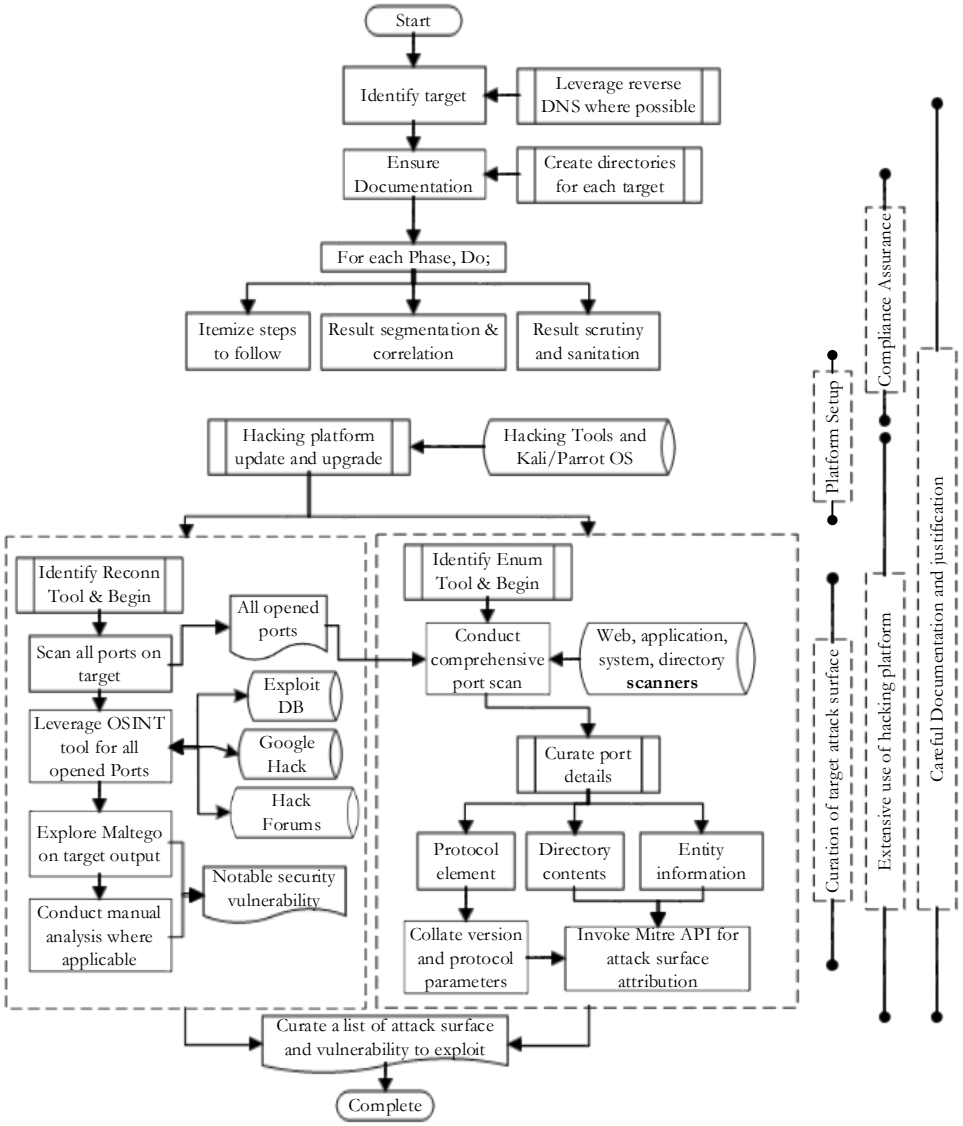


Figure 2: Proposed Reconn and Enum management framework

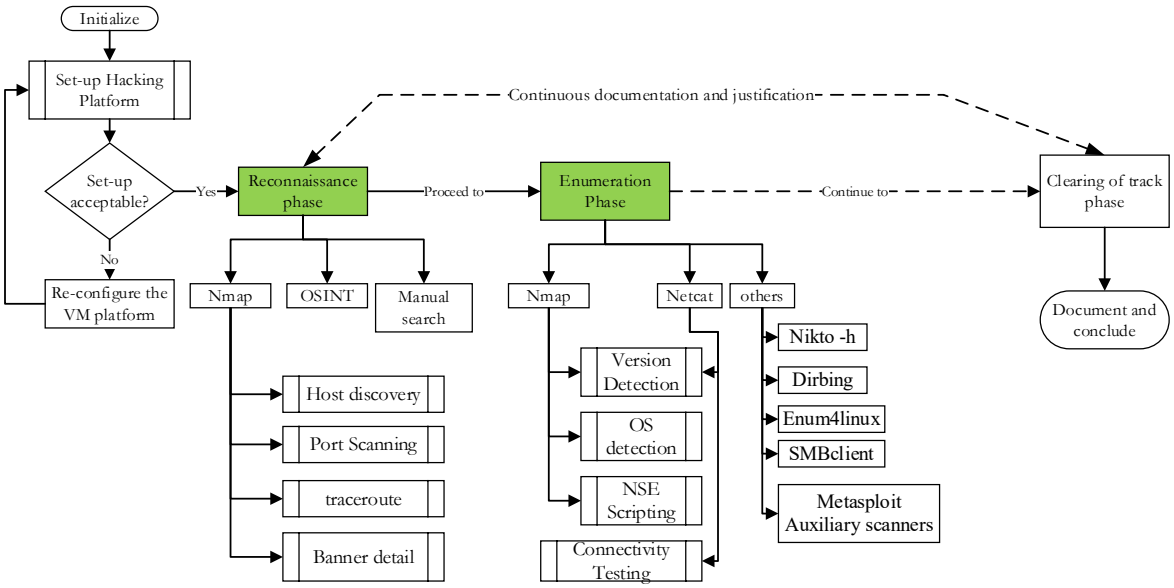


Figure 3: Operational framework for the experimentation process

5. Table 1: Reconnaissance Tools (Nmap)

USAGE	DEPLOYMENT	SUMMARY
Scanning IP Addresses: Command: nmap “ip address”	Linux, Windows, MAC OS X, BSD, Solaris and AmigaOS.	Scans every port on the computer with this IP address.
vulnerability scanning: scripts are in “/usr/share/nmap/scripts” then command: nmap –script= “chosen script” “target”	Linux, Windows, MAC OS X, BSD, Solaris and AmigaOS.	An open-source tool for security auditing and related network discovery. It can be used to identify the current running devices on the systems, hosts, and services available.
Scan multiple targets: Nmap t1,t2,t3..	Linux, Windows, MAC OS X, BSD, Solaris and AmigaOS.	One scan can be done including multiple targets ips at once.
Scan range of hosts: Nmap “range of ip addresses” Nmap 132.123.5.3-20	Linux, Windows, MAC OS X, BSD, Solaris and AmigaOS.	We can scan whole subnets, partial subnets, or file list targets. Nmap can generate possible new targets.
Performing fast scan: nmap -F “target”	Linux, Windows, MAC OS X, BSD, Solaris and AmigaOS.	-F will scan open services, domain names, and ports fast and quickly.
Scan specific ports or entire port ranges: Nmap -p x-xxx localhost Scan specific ports: Nmap -p x,xx x.x.x.x	Linux, Windows, MAC OS X, BSD, Solaris and AmigaOS.	Specifying port can filter the machine running a service on this specific port,
Scan hosts and IP addresses reading from a text file: Nmap -iL xx.txt	Linux, Windows, MAC OS X, BSD, Solaris and AmigaOS.	The “-iL” parameter lets you read from that file, and scan all those hosts for you:
Scan using TCP or UDP protocols: nmap -sT x.x.x.x nmap -sU localhost	Linux, Windows, MAC OS X, BSD, Solaris and AmigaOS.	Sending a UDP packet to each specified port is how the UDP scan operates. Most ports will have an empty packet.
Fragment packets. nmap -f <target IP address> -f split the ip address into tile fragment packets.	Linux, Windows, MAC OS X, BSD, Solaris and AmigaOS.	sending a probe to a network while fragmenting it into numerous smaller packets
Remote Networks:	Linux, Windows, MAC OS X, BSD, Solaris and AmigaOS.	Nmap also displays data about distant networks. In reality, you may use Nmap to analyze a website that you wish to look at, and it will parse the website and find the IP address for that web domain.
If host is vulnerable to DOS: Nmap –script dos -Pn x.x.x.x.	Linux, Windows, MAC OS X, BSD, Solaris and AmigaOS.	A script “NSE” is used to determine if the targeted host is vulnerable to DOS or not.

6. Table 2: Enumeration Tools (Netcat)

USAGE	DEPLOYMENT	SUMMARY
Port scanning: Nc -v -n x.x.x.x x-x	*Pre-installed in Linux and MAC OS. Need to be installed in Windows.*	*To discover open doors and weaknesses in a network. The -v gives us a lengthier result.*
nc -lp port {host} {port}	*Pre-installed in Linux and MAC OS. Need to be installed in Windows.*	Listen for incoming connections
Create a tunnel from one local port to another: nc – xxx | nc x.x.x.x x	*Pre-installed in Linux and MAC OS. Need to be installed in Windows.*	*Using encrypted tunnels within the SSH protocol, tunneling is a method of port redirection. Using an SSH connection, two network devices can communicate by tunneling.*
Encrypt data before transferring over the network: openssl enc -des3 -pass pass:password | nc x.x.x.x x	*Pre-installed in Linux and MAC OS. Need to be installed in Windows.*	sensitive data that is to be uploaded to the cloud should be encrypted on-premises, before upload.

USAGE	DEPLOYMENT	SUMMARY
TCP server mode: -l Nc -l -p x	*Pre-installed in Linux and MAC OS. Need to be installed in Windows.*	Listens for network connections and establishes network connections.
-k Nc -k -l xx	*Pre-installed in Linux and MAC OS. Need to be installed in Windows.*	Listen to the port & IP address after the connection closes
Displaying the IP routing table: Netstat -r	*Pre-installed in Linux and MAC OS. Need to be installed in Windows.*	*The -r displays the IP routing table. -s shows the protocol statistics for the UDP, TCP, SCTP, ICMP, and IP protocols.*

7. Result

Following the operational framework presented in Figure 3, this study explored the application of the proposed framework for an enhanced hacking experience. The testing was carried out on the identified target. A synopsis of the outcome is presented in this section.

a. The Reconnaisance Phase:

The reconnaissance phase is the initial and longest step of a hacking attack. This phase is essentially the preparatory phase in which useful details about a target are gathered; information such as their network, their open and active hosts, and the people involved. It is imperative to note that there are two types of reconnaissance: Active: **Directly** interacting with the target to collect information (Nmap), and Passive: Collecting and gathering information from the target **indirectly**. Common off-the-shelf tools include Nmap and Netcat (the Swiss army knife for network communication). A description of the respective terms used in this phase is provided in Tables 3 and 4. Whilst Table 3 addresses Nmap, Table 4 provides further clarity for Netcat.

7.1.1 Table 3: Description of Nmap Usage

Usage Perspective	Description
Host Discovery + Status	The main objective is to scan and discover which hosts are active and are worth conducting a deeper investigation.
Port Scanning	This forms one of the core operations of the Nmap tool. An attacker can send probes (both normal and carefully crafted probes) to determine if ports are open, closed, or filtered.
Version Detection	If a port is open, Nmap helps determine what version each software/application is running.
OS Detection	Allows attackers to determine and pinpoint the running OS which is extremely helpful as different OSs implement different network standards.
Traceroute	Aids in the finding of network routes
NSE Script Scan	Performs the tasks of detecting service vulnerabilities, gathering a greater amount of information, advanced version detections, as well as malware discovery.

7.1.2 Table 4: Description of Netcat usage

Usage Perspective	Description
Test Connectivity	Allows attacker to gain a shell or reverse connection to the target machine. Additionally, it helps establish numerous connections or backdoors simultaneously.
Port Scanning	Determines the range of ports as well as what ports are up and active.
Version Detection	Helps determine target information such as their running service version.
Banner Details Detection	Hunts down the target's banner details that can be used to exploit vulnerabilities.

Search Exploits	Websites and databases like “Exploit Database” allow attackers to search through any exploit for the specific versions they are looking for.

b. The Enermeration Phase:

The enumeration phase is the second phase of a hacking attack. This phase is extremely critical as it allows attackers to create active connections to the target to perform directed queries to gain additional information. In other words, it allows them to use the information to further enumerate the target to discover their vulnerabilities and weaknesses. Like the reconnaissance phase, the enumeration phase uses tools such as Nmap and Netcat. It is imperative to mention that the enumeration phase follows the same steps as the reconnaissance however, Nmap can be used for various enumeration techniques such as:

1. Nmap for NetBIOS Enumeration: This stands for Network Basic Input/Output System which essentially allows a connection and communication like sharing files to occur over a LAN.
2. Nmap for SNMP Enumeration: This stands for Simple Network Management Protocol in which Nmap gives attackers the ability to use scripts to enumerate and exploit targets.
3. Nmap for LDAP Enumeration: This stands for Lightweight Directory Access Protocol which is used to access directory listings.
4. Nmap for NTP Enumeration: This stands for Network Time Protocol, which is an enumeration process where attackers can pinpoint NTP servers on a network that can be used to implement further enumeration.

The typical output of the enumeration phase includes SNMP data, network shares, IP address table details, a list of password policies, as well as usernames on different systems. For completeness, the other phases of the hacking lifecycle are briefly explained.

c. The Gaining Access Phase: In this phase of a hacking attack, an attacker is granted the opportunity to initiate a sequence of attacks to gain access to the target’s machine and system by relying on the utilization of various tools, methods, and the gathered information regarding the target.

d. The Maintaining Access Phase: This phase of an attack is crucial to an attacker as it determines if they can linger in the target’s system long enough to successfully acquire all the information needed to carry out the attack.

e. The Clearing Tracks Phase: In the final phase of the Ethical Hacking Process, all traces and evidence of the attack are deleted and erased (Ushmani 2018; Yash et al. 2022).

A practical implementation of Nmap and Netcat is further covered in this section. Attackers use these tools to identify and test weaknesses/vulnerabilities in the target networks/web application. As an attacker or penetration tester, several tools and techniques are explored, and some entities prefer to use their speciated crafted private collection of tools for this purpose. A concise depiction of Nmap and Netcat usage is further provided in Table 5. Furthermore, the list of commands used during the process is further provided in Table 6.

Table 5: Summary of Nmap and Netcat command usage

NMAP Command	Description
nmap –sP 192.168.170.0/24	Nmap scan to perform Host discovery
nmap –sn192.168.170.16	Check target host availability with the (-sn) flag specified; this will disable port scanning.
Nmap –Pn –T4 192.168.170.16	Nmap command to perform regular or default scan to check top 100 ports, with (-T4) flag specified to impact the time and speed.
Nmap –Pn –T4 –p- 192.168.170.16	Nmap command scans all 65535 ports with (-p-) flag. Additionally, the (-Pn) flag will disable the host discovery and consider it active to improve scan speed.
Nmap –p80 –O 192.168.170.16	Nmap command to check the target machine's operating system with the (-O) flag. Additionally, I have added (-p80) HTTP port because OS footprinting is impossible without port scanning.

NMAP Command	Description
Nmap –Pn –T4 –p- -sV 192.168.170.16	Nmap command to scan all 65535 ports along with service versions, the (-sV) flag will detect the service version details.
Nmap –Pn –T4 -sV –sC –p 80,110,30109 192.168.170.16	Nmap command to scan to test NSE scripts for (80,110,30109) with (-sC) flag.
NETCAT	
Netcat 192.168.170.16 22	Netcat command to connect to a host on port 22.
Netcat –zv 192.168.170.16 22 Netcat –zv 192.168.170.16 80 Netcat –zvn 192.168.170.16 80	Netcat command with (-z) and (-v) flag performs port scanning and verbose results. Additionally, the (-n) flag prevents DNS resolutions and improves the scan speed.
Netcat –zvn 192.168.170.16 1-100	Netcat command to perform a port scan for the top 100 ports.
Netcat –zvn 192.168.170.16 1-65535	Netcat command to perform a full port scan for all 65535 ports.
Netcat –vn 192.168.170.16 80	Netcat command to grab the banner or version details of the web server.

During the enumeration phase, the attacker establishes an active connection to the system and launches directed queries to learn more about the target. The purpose of obtaining this information is to identify the vulnerabilities of the system of the target to exploit it and perform password attacks to gain illegal access to resources. Therefore, this type of enumeration could be a useful technique for attackers because the target cannot completely avoid being screened by the DNS. By applying the steps provided in Figure 2, the authors were able to extract a complete component of the target system. This includes the use of directory burster tools (dirb), and enhanced enumeration tools (nikto and enum4linux).

7.1.3 Table 6: List of All Commands Used During the Enumeration Phase

Command	Description
ip address (ip a)	display network information and IP address of a machine
netdiscover -r 10.0.2.5/24	Search for the connected machines/host on a specific range of subnets
nmap -sT 10.0.2.5/24	Display the TCP-connected hosts and their open ports
nc -zv 10.0.2.10 1-40000	Netcat command that scans the open ports on the victim machine without sending data, specifies searching area from 1 to 40000
nmap -Pn -sV -A -O 10.0.2.10	-Pn: ping without establishing a connection, -sV: service version, -A -O: discover information about the OS of the host
la /usr/share/nmap/scripts	List all Nmap/NSE scripts
nmap --script=nbstat.nse 10.0.2.10 nmap -sV -v -Pn --script=nbstat.nse 10.0.2.10	*Retrieve the target's NetBIOS names and MAC address*
nmap -sU -p137 --script=nbstat.nse 10.0.2.10 nmap -sU -p139 --script=nbstat.nse 10.0.2.10	sends a UDP/TCP probe on ports 137 and 139 to discover the name of the machine.
nmap -sU -p161 10.0.2.10 nmap -sU -p162 10.0.2.10 nmap -sU -p161 --script=snmp-brute.nse 10.0.2.10	Use brute force guessing to find an SNMP community string on UDP port 161
Nmap -p389 --script=ldap-brute.nse 10.0.2.10 nmap -p389 --script=ldap-search 10.0.2.10	LDAP authentication and searching attempts using brute force
nmap 10.0.2.10 -sU -Pn -p123 --script=ntp-info	uses an NTP server to obtain the time and configuration information
nmap -p25 --script=smtp-enum-users 10.0.2.10	Use the VRFY, EXPN, or RCPT TO commands to list all the users on an SMTP serv,
nmap -sSU -p53 --script=dns-nsec-enum 10.0.2.10	list domain names obtained from the DNS server

8. Discussion

Penetration testing is a crucial aspect of ethical hacking that involves simulating a cyber-attack on a computer system to identify any exploitable vulnerabilities. The testing process involves two key phases: reconnaissance and enumeration. In reconnaissance, information about the target system is gathered passively to inform the testing plan, while in enumeration, detailed information about the target's network resources, services, and usernames is extracted (A. S. B. Singh, Yusof, and Nathan 2021). This study used Nmap for reconnaissance and both Nmap and Netcat for enumeration. As part of the ethical hacking lifecycle, the goal of this report was to highlight the techniques used in the reconnaissance and enumeration phases of penetration testing. In these phases, access to sensitive information and common vulnerabilities can be gained. Netcat, as described, establishes a connection between two computers to transmit data across TCP and UDP transport layer protocols. It also enables banner grabbing to determine the operating system, service, and version on a specific port. On the other hand, Nmap is a versatile tool for network scanning and auditing that can identify security vulnerabilities, detect connected hosts, and determine open ports and running services on the target host and operating system. As highlighted in Figure 3, the exploration of Metasploit auxiliary modules for enumeration is a commonly misconstrued approach. This module can be leveraged like the Nmap, to conduct carefully crafted enumeration.

The ethical hacking lifecycle framework presented in this report outlines the steps involved in the reconnaissance and enumeration phases. It highlights the importance of identifying the target, documenting the process, and utilizing tools such as Kali Linux and Parrot. The report is a valuable resource for beginner hackers, students, and those looking to expand their knowledge in hacking and penetration testing, as it provides detailed explanations of the commands used and how they work. By leveraging this framework, limitations, and constraints often associated with the hacking processes can be easily remedied. This further supports the assertion postulated by Bellaby (2023) which posits that the development of a methodical process could help alleviate the frequent limitations associated with the hacking lifecycle. Ultimately, the goal of ethical hacking is to evaluate the security measures of an organization and potentially gain access to sensitive information. To achieve this, a certain level of trust must be established between the ethical hacker and the business hiring them to perform the test. The framework in this report serves as a guide for conducting the reconnaissance and enumeration phases of penetration testing in a compliant and ethical manner. Arguably, the use of other tools and techniques is also encouraged in this wise. Several off-the-shelf tools and tailored tools have been developed for this purpose. It is therefore worth mentioning that the process of conducting both active and passive information gathering is not limited to the presented framework in Figure 2 and the methodology identified in Figure 3. However, these steps provide a common baseline and a starting point for the same. This baseline aligns with assertions from other related frameworks (Zawali et al. 2021; Ellison, Venter, and Ikuesan 2017; Ellison, Ikuesan, and Venter 2019; Lagrasse et al. 2020; Patil et al. 2017).

9. Related works

Studies on existing ethical hacking frameworks are presented in this section, to highlight the contribution of this current study within the common body of knowledge. Findings in Rafay Baloch's Ethical Hacking and Penetration Testing Guide is a comprehensive book on ethical hacking and penetration testing. The book covers the fundamentals of ethical hacking, such as the different types of attacks, methods, tools, and techniques used in penetration testing. Furthermore, it provides readers with step-by-step instructions and examples to assist them in conducting ethical hacking and penetration testing. The book also emphasizes the significance of adhering to ethical guidelines and safeguarding user data (Baloch 2018). However, no specific detail is provided on the first two phases of the hacking lifecycle. Also, the work failed to provide a relevant methodical framework for conducting these phases. In a dissimilar approach, the study (Bellaby 2023) explored various penetration testing services with a specific focus on penetration testing service limitations. The authors surveyed penetration testing service providers and analyzed the results to determine these services' limitations. The identified limitations include a lack of customization, a limited scope, and difficulty testing real-world scenarios. The study also emphasizes the importance of taking these limitations into account when choosing a penetration testing service.

A study by Hawamleh et al. (2020) further highlights the Importance of Protecting User Data by applying the ethical hacking lifecycle. In this regard, an overview of the different types of hacking attacks, the methods used by hackers, and the consequences of these attacks. The paper also emphasizes the importance of protecting user data and the role of ethical hacking in cyberattack prevention. This source provides an excellent overview of the subject and emphasizes the importance of cybersecurity and ethical hacking. A similar work detailed by

Bhawesh (2022) discusses the different types of hacking attacks, such as SQL injection, cross-site scripting, and man-in-the-middle attacks, and describes the methods used to carry out these attacks. Mitigation strategies against these attacks, such as implementing security software and following best practices for web development are further provided. However, it fails to provide a directed method for conducting ethical hacking, particularly on the modalities for the first two phases. In attempting to identify challenges associated with conducting ethical hacking, a study conducted by Hartley (2015) discusses challenges, such as the increasing sophistication of hacking techniques, the difficulty of detecting and preventing cyberattacks, and the importance of user data protection. Whilst the study is good for information on cyber-attack, it fails to provide any direction or guide on the actual process of ethical hacking. Similarly, a review of cybersecurity risk management models, methods, and frameworks from a penetration testing perspective is presented by the study presented in Shah and Mehtre (2015). The authors examined the existing literature on cybersecurity risk management and identified the key components of effective risk management strategies. Based on the review of the extant studies, the authors discuss risk assessment, risk mitigation, and risk communication, as well as the importance of taking organizational culture and user behavior into account when implementing a risk management strategy.

These existing studies examined and discussed various methods and techniques that can be leveraged during the reconnaissance and enumeration phases of the penetration testing cycle. This framework, on the other hand, provides a more comprehensive and methodical approach to carrying out these phases, making it a valuable resource for readers. The framework provides a systematic and structured approach to reconnaissance and enumeration, lowering the possibility of missing information and increasing process efficiency. This academic paper also emphasizes the significance of understanding the target system and its network environment before proceeding to the next phase to ensure a thorough and complete penetration testing process. Other sources, by contrast, focus on specific techniques or tools and do not provide a comprehensive framework for conducting reconnaissance and enumeration. Overall, this framework is an important addition to the field of penetration testing because it provides a step-by-step guide to conducting effective reconnaissance and enumeration.

10. Conclusion

This study introduced a comprehensive methodology for conducting penetration testing, specifically focused on the reconnaissance and enumeration phases. These initial stages of the ethical hacking process aim to gather information about the target through passive reconnaissance methods and active enumeration techniques that exploit weaknesses and vulnerabilities in direct client-server communication. The proposed framework emphasizes compliance with ethical hacking principles and utilizes various hacking platforms. To assist new learners, users, and hackers, detailed tables (Tables 1, 2, 5, and 6) of Nmap and Netcat usage, deployment, and description are also provided as a clear and concise reference guide. As a future work, the authors intend to further develop a holistic framework that can adapt to any perspective of the hacking lifecycle with clearly stated detail and concepts. Furthermore, such a framework would provide a comprehensive view of the hacking process on which any stakeholder can reliably formulate a penetration testing process. Having such a framework can be a breakthrough for the ethical hacking community, as it would provide a reference model for ethical hacking. To date, the hacking community lacks such.

References

Baloch, Rafay. 2018. “Ethical Hacking and Penetration Testing Guide.” *International Journal of Advance Research in Computer Science and Management* 4 (4): 2253--2257.

Bellaby, Ross W. 2023. “An Ethical Framework for Hacking Operations.” In *The Ethics of Hacking*, 32–52. Bristol University Press.

Bhawesh, Kumawat. 2022. “Ethical Hacking Attacks, Methods, Techniques and Their Protection Measures.” *International Journal of Advance Research in Computer Science and Management* 4 (4): 2253–57. https://madhavuniversity.edu.in/ethical-hacking.html.

Craigen, Dan, Nadia Diakun-Thibault, and Randy Purse. 2014. “Defining Cybersecurity.” *Technology Innovation Management Review* 4 (10).

Ellison, Dagney, Richard Adeyemi Ikuesan, and Hein S. Venter. 2019. “Ontology for Reactive Techniques in Digital Forensics.” *2019 IEEE Conference on Application, Information and Network Security, AINS 2019*, 83–88. https://doi.org/10.1109/AINS47559.2019.8968696.

Ellison, Dagney, Hein Venter, and Adeyemi Ikuesan. 2017. “An Improved Ontology for Knowledge Management in Security and Digital Forensics.” In *European Conference on Cyber Warfare and Security*, 725--733. Academic Conferences International Limited.

GeeksforGeeks. 2022. “History of Cyber Security.” Computer Networks. 2022. https://www.geeksforgeeks.org/history-of-cyber-security/.

Hartley, Regina D. 2015. “Ethical Hacking Pedagogy: An Analysis and Overview of Teaching Students to Hack.” *Journal of International Technology and Information Management* 24 (4): 6.

Hawamleh, A M A, Almuhannad Sulaiman M Alorfi, Jassim Ahmad Al-Gasawneh, and Ghada Al-Rawashdeh. 2020. “Cyber Security and Ethical Hacking: The Importance of Protecting User Data.” *Solid State Technology* 63 (5): 7894–99.

Kebande, Victor R., Nickson M. Karie, and Richard A. Ikuesan. 2020. “Real-Time Monitoring as a Supplementary Security Component of Vigilantism in Modern Network Environments.” *International Journal of Information Technology (Singapore)*. https://doi.org/10.1007/s41870-020-00585-8.

Lagrasse, Maxime, Avinash Singh, Howard Munkhondya, Adeyemi Ikuesan, and Hein Venter. 2020. “Digital Forensic Readiness Framework for Software-Defined Networks Using a Trigger-Based Collection Mechanism.” In *Proceedings of the 15th International Conference on Cyber Warfare and Security, ICCWS 2020*, 296–305. https://doi.org/10.34190/ICCWS.20.045.

Patil, Sonali, Ankur Jangra, Mandar Bhale, Akshay Raina, and Pratik Kulkarni. 2017. “Ethical Hacking: The Need for Cyber Security.” In *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 1602–6. https://doi.org/10.1109/ICPCSI.2017.8391982.

Shah, Sugandh, and Babu M Mehtre. 2015. “An Overview of Vulnerability Assessment and Penetration Testing Techniques.” *Journal of Computer Virology and Hacking Techniques* 11: 27–49.

Singh, Ammrish Singh Beker, Yusnita Yusof, and Yogeswaran Nathan. 2021. “EAGLE: GUI-Based Penetration Testing Tool for Scanning and Enumeration.” In *2021 14th International Conference on Developments in ESystems Engineering (DeSE)*, 97–101.

Singh, Avinash, Adeyemi Ikuesan, and Hein Venter. 2019. “A Context-Aware Trigger Mechanism for Ransomware Forensics.” In *14th International Conference on Cyber Warfare and Security, ICCWS 2019*, 629–38.

Ushmani, Azhar. 2018. “Ethical Hacking.” *International Journal of Information Technology (IJIT)* 4 (6).

Yaacoub, Jean-Paul A, Hassan N Noura, Ola Salman, and Ali Chehab. 2021. “A Survey on Ethical Hacking: Issues and Challenges.” *ArXiv Preprint ArXiv:2103.15072*.

Yash, Tathagat, Suresh Kumar, Kamlesh Sharma, and others. 2022. “Ethical Hacking: A Technique to Enhance Information Security.” In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, 1:780–84.

Zawali, Bako, Richard A. Ikuesan, Victor R. Kebande, Steven Furnell, and Arafat A-Dhaqm. 2021. “Realising a Push Button Modality for Video-Based Forensics.” *Infrastructures* 6 (4): 54. https://doi.org/10.3390/infrastructures6040054.

A Commentary and Exploration of Maritime Applications of Biosecurity and Cybersecurity Intersections

Michaela Barnett[1], Issah Samori[2], Brandon Griffin[3], Xavier-Lewis Palmer[1-4] and Lucas Potter[4]
[1]Blacks In Cybersecurity Headquarters, Inc., VA, USA
[2]Minohealth AI Labs, Accra, Ghana
[3]CySecSol, Franklin, VA, USA
[4]BiosView, Oswego, KS, USA

michaela@bichq.org

Abstract: Prior work has discussed the emerging fields of Biocybersecurity (BCS) and Cyberbiosecurity (CBS) in multiple forms. These include the definition, mission-awareness, general applications, and policy (Murch et al, 2018; Peccoud et al, 2019; Potter et al, 2020). One area that has received relatively little attention are unique BCS/CBS vulnerabilities with maritime theaters, which refers to ocean and littoral-based commercial and military ventures. There is considerable ground for both bioeconomies and militaries to be placed at risk of degraded capacity for activity due to maritime-specific BCS/CBS attacks presently in the future. This is especially the case where aforementioned vulnerabilities are used to disrupt logistics through targeting of personnel and means of transport. This paper discusses the growing relevance of CBS/BCS in maritime space, aspects of maritime environments that can be exploited for BCS attacks, possible BCS/CBS attacks in the near future, possible BCS/CBS means of defense and pre-emptive positioning, and discussion of BCS/CBS relevance in international policy, and differences in application. This paper aims to facilitate and accelerate discussion of BCS to spur helpful action in this area.

Keywords: Biocybersecurity, Cyberbiosecurity Maritime, Biosecurity, Ecology, Military

1. Introduction

War and criminal activities change over time. This is especially true considering that new technologies find their way into offensive actions taken against other parties. Humanity has used sticks and stones from thousands of years back only to investigate new means today through artificial intelligence enhanced malware and remote attacks through DNA in recent times (Ney 2017; Chen et al, 2022). This novelty falls under the scopes of Biocybersecurity (BCS) and Cyberbiosecurity (CBS), which constitute two associated fields at the intersection of biological security and cybersecurity take into account novel biodigital attacks that lean on biospecific knowledge and digital systems (Duncan et al, 2019; Murch et al, 2018; Peccoud et al, 2019; Potter et al, 2020). Prior work by Potter, Powell, Palmer has covered the use of cybercrime and the escalation of such as corollaries of war (2021). In the most basic respect, logistics and supply chains are required for the basic operation of a nation state - and the disruption of maritime shipping through cyberattacks has been very well explored (Crosignani et al, 2023). One specific field of interest to this extrapolation is maritime shipping, which has been facing an increased level of cyberattacks for several years (*Report: Maritime Cyberattacks up by 400 percent*). “seven of the world’s top ten container carriers have publicly acknowledged they have been victims of cyber attacks, with all of the leading four carriers being among the victims” (Chubb, 2022). One particular cyberattack caused more than 300 million USD in losses (and frustratingly the victimized company has not been forthcoming with details of the attack in order to prevent similar ones - all interviewed only spoke on the condition of anonymity) (Greenberg, 2018; Lord, n.d.). The types of cyberattacks on modern maritime industry tend to align with three specific types - “traditional computer systems, industrial control systems and the communications protocols that control the most critical functions on a ship.” (*Are you mitigating maritime cybersecurity risks? | Industryweek*). These systems may include IoT (Internet of Things) technologies in addition to the biological assets that are at risk or are acting as an interlock within bio-digital authentication interfaces (Mileski et al, 2018). A short review of the specific cyber-enabled tools has previously been completed here. One of the most basic tools required for modern shipping AIS (Automatic Identification System) could conceivably be used to remotely interact with air-gapped systems on a modern shipping vessel (Leite Junior et al, 2021). At time of writing, an ongoing cyberattack was being conducted on a Norway-based shipping firm (Treloar, 2023). The physical security of maritime systems has been noted as a vulnerability as the physical security of a cybersystem can often be related to BioCyberSecurity, so we require a review of maritime affairs through the lens of BCS (Tam and Jones, 2018).

To that end, the following paper explores the following topics in series. Foremost, we present the scope of the issue, including a brief summary of the bioeconomy of the US and the globe (section 2). In the next section (section 3) we detail the prime BCS and CBS) issues. Section 4 further evaluates the scale of shipping to the

international bioeconomy. Section 5 explores the threat landscape of the modern maritime environment. Section 6 dives deeper into the autonomous shipping threats that might be prescient in the coming years. Section 7 concludes with an evaluation of ontological security in the context of shipping. Section 8 extends this discussion into policy considerations that may be relevant for future cybersecurity advances. Section 9 details limitations of this study, before concluding. This paper serves as both an exploration and commentary. It is the belief of the authors that there is much depth to be explored in the combination of maritime activities and BCS/CBS policy. Namely, with the rise of ubiquitous computing services and the lowering cost point of cyber-enabled technologies, the goods that we receive for our day-to-day lives will become more at threat than ever: both from actively malicious actors and from improperly vetted measures of automation.

2. Review of the International Bioeconomy vs the United States Bioeconomy

The bioeconomy, which refers to the production of goods and services from biological resources, is heavily dependent on sea-based logistics for the transportation of raw materials, intermediate goods, and finished products. The bioeconomy includes sectors such as agriculture, forestry, fisheries, and food and beverage production, as well as biotechnology and bioenergy. These sectors rely on the transportation of goods by ship to move products to domestic and international markets. According to the United Nations Conference on Trade and Development (UNCTAD) estimates, around 80% of global trade by volume and over 70% by value is carried by sea, therefore, sea-based logistics play a significant role in the bioeconomy (Piñeiro et al, 2021). For example, in the agriculture sector, a large proportion of food exports are transported by sea, and the forestry sector also relies heavily on sea transport for the export of wood products (Rodriguez Franco, 2020). In the fishing industry, much of the catch is transported by sea to processing and distribution centers. The bioenergy sector also relies on sea-based logistics for the transport of biomass and biofuels. In summary, sea-based logistics plays a crucial role in the bioeconomy, as it enables the transportation of raw materials, intermediate goods, and finished products to domestic and international markets, which is essential for the growth and development of the bioeconomy sectors.

In the United States, a significant portion of the bioeconomy is dependent on sea-based logistics, particularly for the export of goods. Considerable exports, such as soybeans, corn, and wheat, are transported by sea to foreign markets (Xavier and Reis, 2020; Gerval, 2022; 김찬우, 2020. The U.S. is also a leading exporter of forest products, practical for export by sea (Rodriguez Franco, 2022). In the fishing industry, much of the catch is transported by sea to processing and distribution centers. In the bioenergy sector, the US is a leading producer of biofuels such as ethanol and biodiesel, and these products are also transported by sea to domestic and international markets. Overall, sea-based logistics plays a significant role in the US bioeconomy, enabling the transportation of raw materials, intermediate goods, and finished products to domestic and international markets, which is essential for the growth and development of the bioeconomy sectors in the US.

3. Cyberbiosecurity & Biocybersecurity Considerations

Cyberbiosecurity and biocybersecurity are related fields, but they have slightly different focuses. In terms of mission specific differences, this is further addressed in Potter and Palmer (2021). Cyberbiosecurity is the practice of protecting biological systems, information, and infrastructure from cyber threats. This includes protecting research facilities, laboratories, and other biological systems from cyber attacks, as well as ensuring the integrity and confidentiality of data generated by biological research. It also includes protecting the critical infrastructure that supports biological research, such as power systems, water supplies, and communication networks. Biocybersecurity, on the other hand, focuses on the intersection of cybersecurity and the life sciences. It deals with the protection of biological systems and organisms from cyber threats, including the protection of biological systems from malicious use, and the protection of organisms from malicious manipulation of their genetic code or other characteristics. Biocybersecurity also includes the protection of the information and data that is generated by biological research, including the protection of intellectual property and proprietary information. In summary, Cyberbiosecurity is focused on protecting biological systems, information, and infrastructure from cyber threats, while Biocybersecurity is focused on the intersection of cybersecurity and the life sciences.

4. Scale of biological imports and relation to Biocybersecurity and Cyberbiosecurity

These features lack an obvious link to BCS. Until the following is considered: an estimated 80% of the world's population relies on imported foods (Dunphy, 2020). In fact, local grown crops can only fulfill the demand for

less than one third of the global population (Kinnunen, 2020). Cost of food importing are projected to reach 2 trillion USD in 2023 (not all of this cost is associated with maritime shipping, but the point remains as to its important link to supplying the world with nutrition) (*Global food imports on track to reach all-time high: FAO | UN News*). Some countries are concerned with the safety of imported food, particularly the US (Center for Food Safety and Applied Nutrition, *FDA strategy for the safety of imported food*).

The unique aspects of the shipping industry that relate to BCS are 1) The use of multiple cyber platforms for various tasks (such as verifying contents, navigating, control of the ship itself, and crew registration), 2) the ability of biologically active components to be shipped, to 3) the exploration and protection of biologically important environments (Grafi, 2022). The first and latter is especially important considering that international trade has led significantly to invasions of unwanted species (Grafi, 2022; Hulme, 2021).

Biocybersecurity refers to the security of biological materials, information, and systems, including measures to protect against the theft, loss, release, or misuse of dangerous pathogens or toxins, as well as the protection of sensitive information related to these materials, from the perspective of biology as an interlock (Potter and Palmer, 2021). Long term goals of those working at this intersection include protecting against accidental releases, as well as intentional ones such as bioterrorism, and also includes protection against cyber attacks on systems and networks used in the handling and transportation of biological materials. Biocybersecurity is an interdisciplinary field that involves collaboration between the biological sciences and information technology to ensure the safety and security of biological materials and information. Biocybersecurity in shipping refers to the security of biological systems, such as those used in cargo or supply chain management, against cyber attacks. These attacks can target the systems used to track and monitor cargo, as well as the systems used to control and operate the ships themselves. Some possible problems that biocybersecurity presents in shipping include: Disruption of cargo tracking and monitoring systems, which can lead to lost or stolen cargo. Tampering with or manipulating the data used to control and operate ships, which can lead to accidents or collisions. The release of harmful biological agents, such as pathogens or toxins, as a result of a cyber attack on cargo systems or ship systems. Overall, biocybersecurity is a growing concern in the shipping industry as it becomes increasingly reliant on digital systems and connected devices. Addressing these risks requires a multi-disciplinary approach that brings together experts in biology, cybersecurity, cyber-physical systems involved and shipping.

5. Threat Landscape of BioCybersecurity in the Maritime Environment

The threat landscape of BioCybersecurity related to the maritime environment includes the exploration, exploitation or manipulation of key systems in navigation, port facilities, maritime infrastructure and their accompanying assets; especially when transporting and maintaining the condition of perishable food or supplies, fuel and maintaining agriculture.

The particularly pertinent attack surfaces here are the information technology systems supporting corporate networks in day-to-day operations through more advanced or crucial systems. This includes navigation and autonomous shipping capabilities. The attacks of these systems commonly involve the use of malicious software to gain unauthorized access to systems and the contained data. In particular, ransomware attacks have become more prevalent (as mentioned in the introduction) and also have combined other elements to increase their effectiveness such as social engineering, that can ultimately lead to the effective phishing of key staff and personnel.

Social engineering attacks can be especially effective in the context of Maritime environments (Guidelines on Cyber Security Onboard Ships, Version Four, n.d.). It can be inferred that threat actors are able to triangulate the element of an individual's emotions and leverage feelings of isolation, home sickness or simply fatigue. Personnel who are responsible for food preparation, water sanitation and any potentially dangerous materials would also be likely to present as a target, as they may overcompensate for any suggested compromise to the communal safety of those aboard.

Ransomware in these systems can be particularly devastating to operational software in maritime environments as they involve encrypting data and demanding payment in exchange for the decryption key, this can delay and augment vital systems. The accompanying phishing technique, typically allows for a malicious link to be triggered by an unknowing end-user to kick off the software used to compromise the user or their system. The delivery medium is usually by SMS (text messages) or email. In 2019, leading into 2020 The United States Coast Guard released a safety alert to inform that a phishing email had led to a facility's network being compromised. This Ryuk Ransomware attack “further burrowed into the industrial control systems that monitor and control cargo transfer and encrypted files critical to process operations" (Gatlan, 2019).

We can infer that the systems that maintain proper condition of transported goods such as refrigeration, agricultural monitoring and managing of exacting substances are at risk and able to be accessed laterally from external facing attack vectors. Maritime Operational Technology has become further ingrained in the Internet of Things (IoT) as well as maintaining infrastructure for independent connections to the wider internet such as; onboard computer systems for personnel and openly available Wifi availability for passengers (Navigating Cybersecurity Challenges in Maritime Operational Technology, 2020).

DDoS attacks involve overwhelming a system or network with traffic, making it unavailable to users. These attacks can be used to disrupt the maritime staff's network or to distract the monitoring security personnel while other attacks are launched. March 2017, in a meeting about the future of Vancouver Energy a DDoS attack occurred on the Port of Vancouver's computer system. This attack occurred as a result of a participant connecting to the port's Wi-Fi, temporarily knocking it offline (Cyber-attacks on the Ports and Maritime Industry – Secolve, n.d.). In evaluating the implications of this not only are end users jeopardizing their methods of daily corporate operation but also have the ability to potentially bridge the unauthorized access allocated by these attacks to more critical systems supporting biological systems and accompanying information.

Supply chain attacks and data breaches can involve the unauthorized access to and theft of sensitive data; involving the compromising of a third-party vendor, integrating technologies or suppliers to gain access to a target system as well as simple mishandling or misappropriation of contributing systems. In the context of Cyberbiosecurity, this might involve compromising the database used to stock and ship requested materials relevant to a nation's bioeconomy or augment their condition mid-transport. Most recently, it was cited that in an evaluation of the factors contributing to the Maritime supply chain by a consulting firm that fault could be assessed in lack of provision of security systems and their configuration versus any percieved absence (Linskey & Palmadesso, 2022). The need for further fortification of Supply chain is also highlighted by the National CounterIntellgence and Security Center (NCSC) and Department of Homeland Security's Cybersecurity and Infrastructure Security Agency (CISA) as well as other agencies by their latest awareness initiative in a comprehensive approach to supply chain reinforcement focusing on the need for Cybersecurity and operational technology comprehsnion within procurement and execution (White House 2023).

Autonomous shipping is constructed to operate either completely autonomously or partially, to aid with certain tasks; utilizing software to manage sensors, cameras and other maritime tools (Wariishi, 2019). This emerging technology is yet another vector for exploitation and a threat to the operational technology hosting and maintaining infrastructure that BioCybersecurity systems rely on in transit.

In reviewing these definitions, scenarios and observations it is important to involve every aspect and resource in this area of security research by exploring the technology used to support the maritime industry and its BioCybersecurity systems. The threat landscape here is evolving with the introduction of Artificial Intelligence and Autonomous Systems that aid in monitoring, reporting and communications which poses a significant risk to critical systems (Navigating Cybersecurity Challenges in Maritime Operational Technology, 2020). Only through a coordinated effort across commercial, recreational and military maritime communities can we move this industry forward and mitigate the impact of threats to operations and commerce.

6. Threats of Autonomous Shipping

Let us explore a concept that is common in many avenues of transportation. The number and quantity of efforts involved in creating autonomous vehicles is growing more numerous. Cyber threats to them have been specified before and must be continually monitored (Akpan et al, 2022; Bolbot et al, 2020; Kavallieratos et al, 2019; Tusher et al, 2022).

There are already several risks associated with automation in maritime logistics, for instance: operational risks, workforce displacement, technology dependence for fundamental tasks, human errors, accidents, and risks of non-compliance with regulatory bodies.

It is important for companies in the maritime logistics industry to be aware of these risks and to implement measures to mitigate them, such as cybersecurity protocols and procedures, regular system maintenance and testing, training of personnel and ensuring that they are compliant with laws and regulations. This problem compounds with biological interlocks or matter within logistics to take into account.

However, the real scale of a compromised autonomous maritime transport system loaded with foodstuffs with a single illicitly stored unit on board which has the ability to autonomously optimize a biothreat - a method previously explored by Potter, Ayala, and Palmer (2021). Such a threat, placed on board an autonomous vessel,

would be able not only to discover the best contaminant to maximize speed of contamination, but theoretically a simple logging or transit change on a compromised logistics unit could inadvertently cause the spread of this disease to other vessels, or even countries.

7. Maritime CBS Applications and Ontological Security

Cyberbiosecurity, maritime applications, and ontological security are interconnected in many ways, particularly in the context of geopolitical tensions and security threats in the South China Sea. Cyberbiosecurity refers to the protection of biological systems and related data from cyber-attacks, while maritime applications involve the use of technology and data in maritime security and navigation. Ontological security refers to the extent to which an individual or group feels secure in their understanding of the world and their place in it. It is a psychological concept that pertains to the way in which people construct their sense of self and their place in the world, and how this sense of self and place is shaped by their interactions with others and the broader social, cultural, and political context in which they live (Griffin et al., 2023). In the context of territorial disputes, ontological security is closely tied to national identity and sovereignty. For example, China's claims in the South China Sea are based on its historical and cultural ties to the region, which are central to its national identity. Taiwan also claims sovereignty over the same islands and reefs, which are a source of national pride and identity.

Maritime law governs activities in the oceans, including territorial disputes between states. The United Nations Convention on the Law of the Sea (UNCLOS) provides a framework for the use and management of ocean resources, including territorial claims, fishing rights, and shipping lanes. The South China Sea, which contains significant oil and gas reserves and important shipping lanes, is the subject of multiple territorial disputes. China claims nearly the entire South China Sea, including islands and reefs also claimed by Taiwan, Vietnam, Malaysia, Brunei, and the Philippines. The United States has contested China's claims and conducted freedom of navigation operations in the region, which has heightened tensions and raised the risk of conflict (United States Department of State Bureau of Oceans and International Environmental and Scientific Affairs Limits in the Seas No. 150 People's Republic of China: Maritime Claims in the South China Sea, 2022).

For example, in 2018, a Chinese cyber espionage group known as TA423 was found to have targeted several companies involved in the South China Sea's energy sector, including the Kasawari gas field and a wind farm in the Taiwan Strait. ("Chinese Hackers Tied to Attacks on South China Sea Energy Firms," 2022). The attack aimed to steal intellectual property and sensitive data related to offshore drilling technology, which could give China a strategic advantage in the disputed region. Such cyber attacks not only threaten the security and competitiveness of individual companies and countries, but also undermine the ontological security of the actors involved, by creating a sense of uncertainty and vulnerability in their technological and strategic capabilities.

In conclusion, the connections between cyberbiosecurity, maritime applications, and ontological security are complex and interdependent, and require a comprehensive and collaborative approach from different stakeholders, including governments, private companies, and international organizations. In the context of the South China Sea dispute, this approach should aim to promote transparency, trust, and cooperation in the use of technology, while also respecting the sovereignty and rights of all parties involved.

8. Policy Considerations, including Defend Forward

There are a number of policies that nations can implement in order to prepare for adequate biocybersecurity. Some examples include: Developing and implementing national biocybersecurity strategies: This would involve creating a comprehensive plan for identifying, assessing, and mitigating biocybersecurity risks. Encouraging public-private partnerships: Governments can work with private sector companies, such as medical device manufacturers and biotech companies, to help identify and mitigate biocybersecurity risks. Investing in research and development: Governments can fund research into new biocybersecurity technologies and techniques in order to stay ahead of emerging threats. Providing education and training: Governments can provide training and education to healthcare providers, researchers, and other stakeholders on how to identify and mitigate biocybersecurity risks. Creating oversight and regulatory bodies: Governments can create oversight and regulatory bodies to ensure that biocybersecurity guidelines and standards are being met by the organizations and companies that operate in these areas. International cooperation: Governments can work with other countries and international organizations to share information and coordinate efforts to address biocybersecurity risks on a global scale. It's important to note that the exact policies will depend on the country and the specific context and that Biocybersecurity policies are multi-disciplinary and will involve multiple actors such as government, private sector, academics, international organizations and so on.

Violations of maritime biocybersecurity or cyberbiosecurity can have significant political ramifications. Here are a few examples: A violation of maritime biocybersecurity could occur if a cyber attacker is able to disrupt the navigation or control systems of a ship, potentially causing a collision or other maritime accident. This could lead to loss of life, environmental damage, and economic disruption, and could result in a diplomatic incident between the countries involved. A violation of cyberbiosecurity in the maritime context could occur if a cyber attacker uses a ship as a platform to launch a cyber attack on other ships or shore-based systems. This could result in a loss of confidence in the safety and security of the maritime industry, and could lead to calls for increased regulation and oversight of ships and shipping companies. A violation of biosecurity in maritime context could occur if a malicious actor uses a ship to smuggle dangerous pathogens into a country. This could lead to a public health crisis and could result in diplomatic tensions between countries. In general, violations of maritime biocybersecurity or cyberbiosecurity can have significant economic, political, and security implications, and can cause damage to the reputation and credibility of the countries and organizations involved. It is important for countries to have robust cyber and biosecurity measures in place to protect against these types of incidents.

Defend Forward is a cyber security strategy framework developed by the U.S. Cyber Command (USCYBERCOM), it is designed to proactively detect and disrupt malicious cyber activities before they can do harm to U.S. interests. This framework focuses on identifying and stopping threats at their source, rather than simply reacting to them once they have already entered the network. Defend Forward could be a suitable framework in which to include maritime applications biocybersecurity as it's a proactive approach to detecting and disrupting malicious cyber activities before they can cause harm to U.S. interests. However, incorporating maritime applications biocybersecurity into the framework would require some additional considerations. Maritime applications biocybersecurity refers to the protection of ships, ports, and other maritime infrastructure from cyber threats that could disrupt or damage these systems. This could include protecting against the hacking of navigation and control systems on ships, or the spread of computer viruses through biological vectors on ships. To include maritime applications biocybersecurity within the Defend Forward framework, the following steps could be taken:

1. Identifying the specific biocybersecurity risks associated with maritime systems
2. Developing and implementing biocybersecurity protocols and guidelines for maritime systems specifically tailored to the maritime industry, that would be in line with the overall Defend Forward strategy
3. Encouraging public-private partnerships between the maritime industry and government to share information and coordinate efforts to address biocybersecurity risks
4. Investing in research and development for maritime biocybersecurity technologies and techniques
5. Providing education and training for maritime industry workers on how to identify and mitigate biocybersecurity risks
6. Creating oversight and regulatory bodies to ensure that biocybersecurity guidelines and standards are being met by the organizations and companies that operate in the maritime industry,

`It's important to note that incorporating maritime applications biocybersecurity into the Defend Forward framework would require a comprehensive approach that involves multiple actors such as government, private sector, academics, international organizations and so on, and that the strategies and protocols need to be reviewed and updated regularly as the field of biocybersecurity is evolving.

9. Limitations

`The limitations of this paper should be acknowledged to provide a comprehensive understanding of the group research. Firstly, the study mainly relied on resources from limited outlets utilized by academic researchers versus those participating actively as full-time practitioners. The findings may not be generalizable to other maritime organizations or sectors. Secondly, the study was conducted during a specific period, and the results may be affected by the changing circumstances of the time. Therefore, the findings may not reflect the current state of the maritime industry in relation to the Cybersecurity implications. Further research is needed to validate these findings and explore other potential limitations.

10. Conclusion

Existence on earth is a fragile thing, as so many know all too well. So too are the tenuous links that connect our daily lives with the goods that we require - from the most basic foods that sustain us to the many medicines that

are required - they mainly arrive via sea through maritime shipping. So we see that through the lens of BCS and CBS, these tenuous links are far too crucial to remain both unexamined and unprotected from modern methods of attack. Even while more advanced and even autonomous methods of shipping are being explored, their attack surfaces must be continually monitored and impacts of cyberattacks to the processes of shipping estimated (Weaver et al, 2022).

Acknowledgments:

The authors would like to express their gratitude to the individuals in the Maritime Cybersecurity industry that have contributed to this research.

References:

Akpan, F., Bendiab, G., Shiaeles, S., Karamperidis, S. and Michaloliakos, M., 2022. Cybersecurity Challenges in the Maritime Sector. *Network*, *2*(1), pp.123-138.

Bolbot, V., Theotokatos, G., Boulougouris, E. and Vassalos, D., 2020. A novel cyber-risk assessment method for ship systems. *Safety science*, *131*, p.104908.

Carballo Piñeiro, L., Mejia Jr, M. Q., & Ballini, F. (2021). Beyond COVID-19: the future of maritime transport. *WMU Journal of Maritime Affairs*, *20*(2), 127-133.

Center for Food Safety and Applied Nutrition (no date) *FDA strategy for the safety of imported food*, *U.S. Food and Drug Administration*. FDA. Available at: https://www.fda.gov/food/importing-food-products-united-states/fda-strategy-safety-imported-food (Accessed: January 22, 2023).

Chen, Z., Liu, J., Shen, Y., Simsek, M., Kantarci, B., Mouftah, H. T., & Djukic, P. (2022). Machine learning-enabled iot security: Open issues and challenges under advanced persistent threats. *ACM Computing Surveys*, *55*(5), 1-37.

Chinese Hackers Tied to Attacks on South China Sea Energy Firms. (2022, August 30). Bloomberg.com. https://www.bloomberg.com/news/articles/2022-08-30/chinese-hackers-tied-to-attacks-on-south-china-sea-energy-firms#xj4y7vzkg

Chubb, N. (2022) *Cyber attacks: Who targets the Maritime Industry and why?*, *Thetius*. Available at: https://thetius.com/cyber-attacks-who-targets-the-maritime-industry-and-why/ (Accessed: January 22, 2023).

Crosignani, M., Macchiavelli, M. and Silva, A.F., 2023. Pirates without borders: The propagation of cyberattacks through firms' supply chains. *Journal of Financial Economics*, *147*(2), pp.432-448.

Dunphy, S. (2020) *Majority of the world's population depends on imported food*, *European Scientist*. Available at: https://www.europeanscientist.com/en/agriculture/majority-of-the-worlds-population-depends-on-imported-food/ (Accessed: January 22, 2023). Tam, K. and Jones, K.D., 2018. Maritime cybersecurity policy: the scope and impact of evolving technology on international shipping. *Journal of Cyber Policy*, *3*(2), pp.147-164.

Gatlan, S. (2019, December 27). U.S. Coast Guard Says Ryuk Ransomware Took Down Maritime Facility.

Gerval, A., 2022. US Agricultural Trade Showed Resiliency Through COVID-19 Pandemic. Amber Waves: The Economics of Food, Farming, Natural Resources, and Rural America, 2020.

Global food imports on track to reach all-time high: FAO | UN News. (2022, November 11). News.un.org. https://news.un.org/en/story/2022/11/1130467

Grafi, A. (2022) Why ports are at risk of cyberattacks, Dark Reading. Available at: https://www.darkreading.com/attacks-breaches/why-ports-are-at-risk-of-cyberattacks (Accessed: January 22, 2023).

Greenberg, A. (2018) The untold story of notpetya, the most devastating cyberattack in history, Wired. Conde Nast. Available at: https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/ (Accessed: January 22, 2023).

Griffin, B., Alexander, K., Palmer, X.-L., & Potter, L. (2023). Social-Engineering, Bio-economies, and Nation-State Ontological Security: A Commentary. International Conference on Cyber Warfare and Security, 18(1), 111–118. https://doi.org/10.34190/iccws.18.1.1021

Guidelines on Cyber Security Onboard Ships, Version Four. (n.d.). Www.ics-Shipping.org. https://www.ics-shipping.org/publication/guidelines-on-cyber-security-onboard-ships-version-four/

Hulme, P.E., 2021. Unwelcome exchange: International trade as a direct and indirect driver of biological invasions worldwide. *One Earth*, *4*(5), pp.666-679.

Are you mitigating maritime cybersecurity risks? | Industryweek (no date). Available at: https://www.industryweek.com/technology-and-iiot/cybersecurity/article/21216344/are-you-mitigating-maritime-cybersecurity-risks (Accessed: January 23, 2023).

Kavallieratos, G., Katsikas, S., Gkioulos, V. (2019). Cyber-Attacks Against the Autonomous Ship. In: , et al. Computer Security. SECPRE CyberICPS 2018 2018. Lecture Notes in Computer Science(), vol 11387. Springer, Cham. https://doi.org/10.1007/978-3-030-12786-2_2

김찬우, 2020. *Environmental Impacts of International Crop Trade on Supply Chain from Farm to Market* (Doctoral dissertation, 서울대학교 대학원).

Kinnunen, P., Guillaume, J.H., Taka, M., D'odorico, P., Siebert, S., Puma, M.J., Jalava, M. and Kummu, M., 2020. Local food crop production can fulfil demand for less than one-third of the population. *Nature Food*, *1*(4), pp.229-237.

Leite Junior, W.C., de Moraes, C.C., de Albuquerque, C.E., Machado, R.C.S. and de Sá, A.O., 2021. A triggering mechanism for cyber-attacks in naval sensors and systems. *Sensors*, *21*(9), p.3195.

Linskey, D., & Palmadesso, C. (2022, April 20). *Maritime Supply Chain Crisis: 6 Ways to Effectively Mitigate Port Security Risks*. Kroll. https://www.kroll.com/en/insights/publications/maritime-supply-chain-crisis-effectively-mitigate-port-security-risks

Lord, N. (no date) *The cost of a malware infection? for Maersk, $300 million*, *Digital Guardian*. Available at: https://digitalguardian.com/blog/cost-malware-infection-maersk-300-million (Accessed: January 22, 2023).

Mileski, J., Clott, C. and Galvao, C.B., 2018. Cyberattacks on ships: a wicked problem approach. *Maritime Business Review*.

Navigating Cybersecurity Challenges in Maritime Operational Technology. (n.d.). The Maritime Executive. https://maritime-executive.com/editorials/navigating-cybersecurity-challenges-in-maritime-operational-technology

Ney, P., Koscher, K., Organick, L., Ceze, L., & Kohno, T. (2017, August). Computer Security, Privacy, and DNA Sequencing: Compromising Computers with Synthesized DNA, Privacy Leaks, and More. In *USENIX security symposium* (Vol. 26, pp. 765-779).

Palmer, X.L., Powell, E. and Potter, L., 2021, June. Biocyberwarfare and Crime: A Juncture of Rethought. In *European Conference on Cyber Warfare and Security* (pp. 517-XIV). Academic Conferences International Limited.

Potter, L., Ayala, O. and Palmer, X.L., 2021, February. Biocybersecurity: A Converging Threat as an Auxiliary to War. In *ICCWS 2021 16th International Conference on Cyber Warfare and Security* (p. 291). Academic Conferences Limited.

Potter, L. and Palmer, X.L., 2021, April. Human Factors in Biocybersecurity Wargames. In Future of Information and Communication Conference (pp. 666-673). Springer, Cham.

Report: Maritime Cyberattacks Up by 400 Percent. (n.d.). The Maritime Executive. https://maritime-executive.com/article/report-maritime-cyberattacks-up-by-400-percent

Rodriguez Franco, C., 2022. Forest biomass potential for wood pellets production in the United States of America for exportation: a review. *Biofuels*, pp.1-12.

Treloar, S. (2023) *Cyber attack hits 1,000 merchant ships as Norway firm targeted*, *Bloomberg.com*. Bloomberg. Available at: https://www.bloomberg.com/news/articles/2023-01-18/cyber-attack-hits-1-000-merchant-ships-as-norway-firm-targeted (Accessed: January 22, 2023).

Tusher, H.M., Munim, Z.H., Notteboom, T.E., Kim, T.E. and Nazir, S., 2022. Cyber security risk assessment in autonomous shipping. *Maritime Economics & Logistics*, pp.1-20.

Xavier, D.L.D.J. and Reis, J.G.M.D., 2022. Social Network Analysis on Agricultural International Trade: A Study on Soybean, Soybean Cake and Maize Exports. *Chemistry Proceedings*, *10*(1), p.37.

White House Office of the National Cyber Director. (2023, April 11). *April is Supply Chain Integrity Month*. The White House. https://www.whitehouse.gov/oncd/briefing-room/2023/04/11/april-is-supply-chain-integrity-month/

Wariishi, K. (2019, September). *Mitsui & Co. Global Strategic Studies Institute Monthly Report*. https://www.mitsui.com/mgssi/en/index.html.

Weaver, G.A., Feddersen, B., Marla, L., Wei, D., Rose, A. and Van Moer, M., 2022. Estimating economic losses from cyber-attacks on shipping ports: An optimization-based approach. *Transportation Research Part C: Emerging Technologies*, *137*, p.103423.

U.S Coast Guard Says Ryuk Ransomware Took Down Maritime Facility. (n.d.). BleepingComputer. Retrieved April 11, 2023, from https://www.bleepingcomputer.com/news/security/us-coast-guard-says-ryuk-ransomware-took-down-maritime-facility/

United States Department of State Bureau of Oceans and International Environmental and Scientific Affairs Limits in the Seas No. 150 People's Republic of China: Maritime Claims in the South China Sea. (2022). https://www.state.gov/wp-content/uploads/2022/01/LIS150-SCS.pdf

Functional Architectural Design of a Digital Forensic Readiness Cybercrime Language as a Service

Stacey Omeleze Baror[1], Richard Adeyemi Ikuesan[2] and Hein S. Venter[3],
DigiForS Research Group, University of Pretoria, South Africa[1,3] and Computing and Applied Technology Department, College of Technological Innovation, Zayed University, UAE[2]

stacey.baror@cs.up.ac.za[1]
richard.ikuesan@zu.ac.ae[2]
hventer@cs.up.ac.za[3]

Abstract: Developing a generic digital forensic solution in a cloud computing platform that can address the functional requirements of digital forensic stakeholders is a complex process. The solution would require a technology-independent architectural design that addresses the challenges of incident threat identification, triggering, incident threat isolation and investigation. Existing approaches are limited to the functionality that treats these four challenges individually without the due diligence to consider their interoperability. This study proposes a context-independent and technology-neutral architecture to address these issues by developing a digital forensic readiness (DFR) based on a human language communication interaction (HLI) system that could create a cybercrime language as a service (DFClaaS). The functional architectural design of the proposed DFR HLI DFClaaS system comprises microservices, layered and event/component-based architectural patterns on top of cloud architectural patterns. The DFR HLI DFClaaS system integrates flexibility and other quality requirements to separate concerns while accommodating rigid requirements like security and reliability. The developed architecture is essential for any human-centred digital forensic solution. Therefore, integrating the developed architecture presents a reliable baseline for the digital forensic community.

Keywords: Cybercrime language, Semantic identifier, NLP, Digital forensic readiness, Cybercrime lexicon.

Introduction

The purpose of digital forensic software applications is to determine the flow of events before, during and after an alleged incident. It identifies the cause of an incident and the events preceding it, proving the consistency of potential evidence recovered. Despite the available software for digital forensic incident discovery and investigations, there is a growing demand for software that can proactively address cybercrime attacks that are non-technical-based (Omeleze and Venter, 2013); (Valjarevic et al. 2014). Therefore, this means that the approach to designing such digital forensic applications must adapt to rapid software system design, development and reusable subsystems.

This study proposes a fundamental component to actualizing this goal by developing a context-independent functional architecture of a digital forensic readiness (DFR) framework based on natural human language communication and interaction (HLI) that uses service to show the various components, such as cybercrime language as a service termed digital forensic readiness cybercrime language as a service (DFCLaaS). Investigation of digital forensic in a cloud could be infeasible due to the volatile nature of cloud instances. One approach often employed to address the anti-forensic characteristic is the software deployment technique as a forensic readiness process. Given the volatile nature of cloud instances, studies have adduced that digital forensic readiness is one of the ways to safeguard cloud computing (Singh et al., 2022).

In a previous study (Baror et al., 2021), the digital forensic readiness (DFR) framework for human-natural language interaction was proposed, albeit without a functional design architecture. For this study, an architectural design employs a technology-neutral software requirement specifications design and the use-case responsibility-driven by analysis and design (URDAD) methodologies. The URDAD is a service-oriented method to construct a domain-specific language (DSL) (Solms et al., 2011). It uses microservices architecture in a logical design of a DF tool. The separation of concerns is achieved at both the abstraction level and deployment phases using model-view-controller architectural patterns. Therefore this study proposed a functional architectural design of a DFR HLI cybercrime language as a service.

This paper proposes a context-independent and technology-neutral functional architectural design of a digital forensic readiness software application using human language interaction. The proposed approach applies to any digital forensic readiness solution, especially for the SaaS public cloud. This study considers a context-independent approach to digital forensic readiness. Furthermore, this study introduced microservice architecture as an independent interface for forensic readiness without loss of investigation interactivity. Again,

this provides a comprehensive taxonomy of stakeholders in the forensic process through a use-case scenario to address different stakeholders' concerns during the tool design and development.

This paper is structured as follows: An introductory overview of a human-natural language interaction system's digital forensic readiness architecture. Then a presentation of a proposed approach and evaluation of the architectural design. A conclusion with proposed future studies.

The Proposed DFR HLI DFClaaS System

In order to improve the efficiency, performance and scalability of any extensive system, serverless computing is one approach. Serverless computing enables cloud services to provide a big data-focused architecture to users. The architectures are made up of various sub-services (i.e., microservices) that facilitate the end-to-end extract, transform and load (ETL) workflow needed to accomplish data analysis tasks (Chen et al., 2013). This study proposed a digital forensic readiness framework that utilizes human-natural language interaction, referred to as the DFClaaS system, that leverages microservices.

The DFClaaS system presents technology-neutral software requirement specifications, starting with the functional design of a digital forensic readiness application. It uses the use-case responsibility-driven analysis and design (URDAD) methodology to achieve the user's needs Solms et al. (2016). This is because the URDAD is a domain-specific software design method. Furthermore, detailed software requirements specifications (SRS) and engineering process is employed to justify the architectural decisions at various steps of the system development. The functional requirements of the DFR HLI DFClaaS system are presented as the first step designs that trigger the architectural decisions.

2.1 The Functional Requirements of DFR HLI DFClaaS System

The functional requirements of a system are presented as a natural, plain and understandable language to all stakeholders. It identifies and defines the set of system components, inputs, the behaviour of the information at processing, and the expected, resulting output (Len Bass Paul Clements 2012). It focuses on the behavioural conditions that describe the use cases by capturing the role of the system's users and applying their various functions to the system (Len Bass Paul Clements, 2012). Functional requirements thus provide an in-depth view of the various subsystems of the DFClaaS.

Table 1: DFR HLI DFClaaS Functional Requirements

ReqNo	Requirements Description	ReqNo	Requirements Description
FR1	The DFR HLI DFClaaS system must enable users to sign up and create an account	FR2	The DFR HLI DFClaaS system must allow authorized users to access resources.
FR3	The DFR HLI DFClaaS system should restrict users' actions and access to data based on their user permissions	FR4	The system should be able to analyze the parsed data to determine anomalies, patterns, and relationships within the data.
FR5	The DFR HLI DFClaaS system should allow DF investigators and other authorized users to input search query criteria	FR6	The system must be able to auto-identify and verify ISP and cloud service provider.
FR7	DFR HLI system should allow the system manager (admin) user to maintain and manage the permission of all users registered to access the system	FR8	The DFR HLI DFClaaS system should be able to analyze the parsed data to determine anomalies, patterns, and relationships within the data.
FR9	The users should be able to view, update and edit submitted cybercrime incident	FR10	The system must validate the date and time of incident report submission
FR11	The users must be able to notice if a user continually uploads a cybercrime reports	FR12	The system must enable users to report cybercrime without the requirement to submit any personal details.
FR13	The DFR HLI DFClaaS System must be able to allow users to report cybercrime	FR14	The System must allow users to describe cybercrime incidents in text, and details of the incidents, such as time, date, and location.
FR15	The system should identify anomalies within the parsed data	FR16	The system should analyze semantics, and sentiments within the parsed data.

ReqNo	Requirements Description	ReqNo	Requirements Description
FR17	The system should identify patterns within the text data.	FR18	The system should determine relationships within the parsed text data semantics and sentiments
FR19	The system should visualize the gathered and analyzed data in multiple ways	FR20	The system should be able to allow users to subscribe to the DFR cybercrime semantic and language identifier.
FR21	The system should display a map showing analyzed and gathered information.	FR22	The system should display a graph link diagram showcasing relationships between the data.
FR23	The system should notify a user if an anomaly occurs.	FR24	The system should visualize the gathered data in a timeline diagram.
FR25	The system should be able to give access to authorized and verified users to download and use the DFR semantic trigger API.	FR26	The system should display a graph link diagram showcasing relationships between the data..
FR27	The system should be assigned roles to users such as the DF investigator to gain access to the system.	FR28	The system should be able to allow authorized independent legal or judiciary teams to verify the validity of test data used in the development of the cybercrime language trigger
FR29	The system must be able to show the processes used to preserve the potential digital evidence.	FR30	The authorized users should be able to upload the forensic reports of past cybercrime incident investigations to the digital forensic cybercrime language database
FR31	The system should provide access to an authorized user to activate and use the DFR semantic trigger API.	FR32	The system should allow users to exit the use of the DFR HLI DFCLaaS application system at any given time.
FR33	The DFR HLI DFClaaS system needs an interface to com- municate with the database	FR34	The DFR HLI DFCLaaS system needs to send an email to the System Admin (manager) that adds cybercrime reports.

The functional requirements (Table 1) should be addressed to design the architecture of any software application system. The users' requirements align with the system's overall requirements, so the properties of the architectural and requirement attributes are adequately selected. The next section presents a conceptual design of the proposed DFR HLI DFClaaS framework.

2.2 Microservices Architectural Design Structure of the DFR HLI System

The architecture at the top-level design is the cloud reference architecture (Len Bass Paul Clements, 2012). The microservice architectural styles are chosen as the core architectural patterns to implement the system, as shown in Figure 1. The individual microservice is further represented as a layered architecture with a model-view-controller architecture style to support the component-to-component and user interaction flexibility, as shown in Figure 2. The system consists of functionality and connections with various parts (see Figure 1). Therefore, using microservice architectural styles achieves loosely coupled. The choice of microservices is such that re-usability and flexibility quality attributes are met without trade-off security. For the DFClaaS system, the microservice architectural style allows for a large application to be subdivided into smaller independent services, whereby each service has its role and responsibility.

Firstly, as depicted in Figure 1, the microservice architecture is built on layered architectural pattern depicted in Figure 2. Then an event/component-based architecture to allow for a loosely coupled system and an independent performance-enhanced system, as shown in Figure 2. The microservice function exposes and consumes functionality using contracts, events, and messages. The component-based architecture does create reusable functional and logical components of the DFR HLI DFClaaS design that are location transparent and expose well-defined communication interfaces. The link between these components and the upper layer is the API. Thus, the API serves as the primary interface gateway for the system. The benefits of a microservice architecture are individual processes that communicate with each other over the network to fulfil a goal.

2.2.1 API Gateway

The API gateway service of the DFR HLI DFClaaS system facilitates the integration. It integrates external services, such as security and other device-specific APIs. The API gateway provides a single entry point for all the client applications to

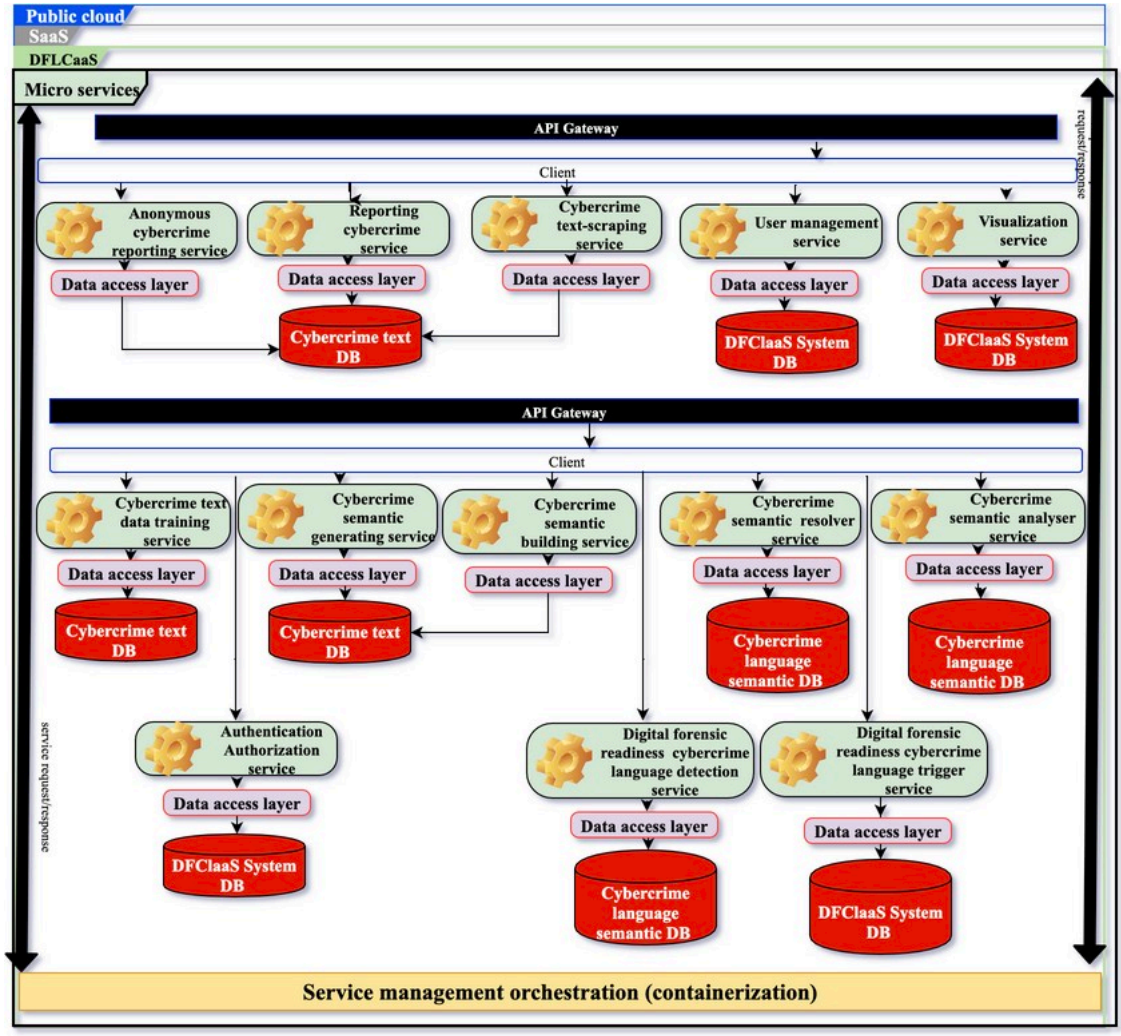


Figure 1: The Microservices Architectural view of the DFR HLI DFClaaS System

communicates with the service requests and receives responses per the request. The API gateway also facilitates communication between the different services. Heavy-tailed calls could quickly slow down the system's performance at the user interface. Web interfaces built on single-page applications (SPA) frameworks such as Angular and React (further explained in future work that presents the prototype implementation of the DFClaaS system) could encounter similar issues. The integration requirements ensure successful deployment criteria following the architectural structure. The DFR HLI DFClaaS system consists of the following services:

2.2.2 Anonymous cybercrime reporting service

The anonymous cybercrime reporting service allows victims of cyberattacks to report the incident without providing any reporter's personal information. They also facilitate the provision of cybercrime strings to increase the volume of cybercrime data. The value of the string data provided by the victims is an input used to build the database that further creates the cybercrime semantic, lexicon and language components of DFClaaS.

2.2.3 Reporting cybercrime service

The reporting cybercrime service of the system is aligned with data gathering of the interactive communication between the cybercrime victim and the cyber attacker. In the process, collect and store data input for building cybercrime language patterns. The reporting cybercrime service gathers data from external sources that can be,

parsed, cleaned and trained using both the natural language processing techniques and the machine learning components (see Figure 1). During data processing, string data parsing determines the sentiment and builds semantics to create a DFR cybercrime language. This is achieved by running the parsed string data through the deep learning model to measure the sentimental, numeral and contextual differences between the reported and stored historical cybercrime string data in the process identifies anomalies.

2.2.4 Cybercrime text scraping service

The cybercrime text scraping service is centred on allowing the system to carry out the function of web data extraction, scrapping and extracting string data from web pages by accessing the World Wide Web using the web browser. It is one of the means to increase the volume of the cybercrime data set available for training. The cybercrime text-scraping service automates the processes of using a bot or web crawler to gather cybercrime-specific string data.

2.2.5 User management service

The user management service is the human interface interacting with the system. It can be a web or mobile app, the application's client side. The model-view-controller of the layered architecture separates the user interaction from other parts of the system that is a microservice that contains the structure of the DFR HLI DFClaaS (Baror et al., 2021) (Baror et al., 2020),(Baror and Venter, 2019). The microservice architecture facilitates all users' actions, and layered architecture manages activities such as logging in, creating and updating profile details to access the system, and uploading, viewing or accessing the DFR HLI application system. The layered architecture comprises the access, business logic, and persistence layers. The access layer allows communication between the user management service through a controller, while the business layer contains all the logic for the user management service. The persistence layer involves storing, retrieving, updating, and the atomicity, consistency, isolation, and durability (ACID) functions. The ACID properties of the database transactions ensure that the cybercrime data used to develop the DFClaaS are valid.

2.2.6 Visualization service

The data visualization service provides the functionality that allows authorized users to visualize raw data, analyze it, identify patterns, and enable the user to use the historical string data to visualize cybercrime language patterns in the DFClaaS system. This service benefits the users; it enhances the usability quality requirement. The model-view-controller architectural patterns, to be discussed in detail in section 2.3.1, are employed to achieve the visualization functions of the system. The view (visualization) is a separate entity from the model controller. Another advantage of having the visualization services of the system is that it allows for easy modifications, interoperability and integration of the visualization component.

2.2.7 Cybercrime text-data training service

The training of the cybercrime text data employs a deep learning auto-encoder technique that narrows the accuracy of the resultant cybercrime string data trigger using recurrent neural networks (RNN). The RNN relies on feature extraction to extract salient features from documents represented using word embedding. The RNN processes sequential temporal information, such as sentences, frequently added to the reported cybercrime database (see Figure 1). The autoencoder deep learning service learns efficient data representation (encoding) by training the network to ignore signal (noise). It is associated with the transformation and extraction of features to establish a relationship between the stimuli, the historical cybercrime text data, and associated neural responses, in the form of a 'new' document or other form of string data. In training the collected cybercrime text data, the propagation of an accurate forensic process is first employed. This implies that the data collected must follow the rule of evidence to maintain the chain of custody and rule of evidence. This is by applying strict access control, maintaining logging, and authorization mechanisms

The central concept of the DFR HLI system is to build an all-in-one digital forensic readiness system that uses reported cybercrime and cybercrime-related data sets to carry out data analysis, feature, semantic and lexicon extractions and to create a digital forensic investigation-ready output that could use the user's language interaction styles as trigger detection.

2.2.8 Cybercrime semantic builder service

The cybercrime semantic builder service uses the input from the generated semantic string. The cybercrime semantic database uses a lexical-semantic analyzer to determine the validity of the available cybercrime string. The semantic information is analyzed for consistency and similarities to the historical cybercrime string. The cybercrime semantic database builder provides the platform to generate semantics uniquely available to forensic investigators.

2.2.9 Cybercrime semantic analyser & resolver service

The cybercrime semantic resolver in this study concerns meanings, words reference (denotation) and associated concepts (connotations). It further focuses on semantic reconciliation to determine the semantics value of the historical cybercrime text data. In this case, the source is meant to compare incoming text data to the existing database of the receiving API in the language service. The function is to create a valid semantic to construct a digital forensic cybercrime language trigger service. In NLP, the two main components of semantics are logical and lexical semantics (Hirschberg and Manning, 2015),(Zellers et al., 2019), which are the baseline for developing the DF cybercrime language. The cybercrime semantic generating, building and resolving services of the system use the cybercrime text data for lexical semantic functions, which are some of the processes employed to create the DFClaaS service. The authors (Baror et al., 2021) summarised the lexical semantics that analyzes meanings of the cybercrime text that have been gathered from morphology similarity, sentiment and the relationship between the words. The cybercrime semantics are generated in two instances, from the cloud user's platform, for example, the SMEs business owners and the reported cybercrime database service (see Figure 1). The analyser extracts sentiments, language patterns and semantics to identify potential trends within a cybercrime corpus. The corpus are further explored with deep learning autoencoder, recurrent neural networks, feature extraction and clustering deep learning techniques to continually analyze the 'newly' submitted and existing text data stored in the historically reported cybercrime text data Figure 1. The cybercrime semantic resolver compares the instances and selects the appropriate option as output. In cases where instances are not resolvable, the corresponding data is further parsed over to the text-data training service and semantic builder service for consideration.

2.2.10 Digital forensic readiness cybercrime language detection and trigger service

The digital forensic readiness cybercrime language detection and trigger services are concerned with the forensic readiness language of the DFR HLI DFClaaS framework. In an earlier research (Baror et al., 2021), the authors described the process employed to extract semantic value from historical cybercrime text data in a SaaS cloud platform. It performs a similarity check between the historical cybercrime text data and the SaaS cloud instance semantic to activate the detection and the trigger service. It combines a cloud-based lexicon and a cybercrime lexicon executed with pattern observation and extraction that assumes a phrase's collective polarity is the sum of member words' polarity in conjunction with the cloud lexicon comparison to extract a potential cybercrime language trigger. For the detection to occur, actions like validation of the semantic similarity of the text data of an 'in-flow' of SaaS cloud instance and the corresponding cybercrime historical text data in the cybercrime database are fed to the training services of the system. Furthermore, the activities that the cybercrime language detection implements in the NLP pipeline, such as the parsing service that provides the functionality of extracting and parsing text data in a format that will be useful for creating cybercrime semantics and the language trigger service. Both contents are fused into the cybercrime semantic database builder to activate the language detection and trigger services.

2.2.11 Authentication & Authorization services

The function of the authentication and authorization services of the DFR HLI DFClaaS system handles two aspects: verifying a user's identity to allow the user access the DFR HLI DFClaaS system. While the authorization functions enable the user permission to access specific resources, the user is supposed to gain access to the functions, components and other resources of the DFR HLI DFClaaS system, ensuring that users are who they claim to be. Furthermore, an authorized user's access is created using specific user details such as identity number, name, surname, email, and a random password. An authorized user's identity (ID) information is static and constitutes the primary key in the ACID domain of the system. As indicated in section 2.2, the core architectural style of the DFR HLI system is the microservice architecture. The following section describes how the architectural styles are applied to the DFR HLI system.

2.3 Layered Architectural Patterns of the DFR HLI DFClaaS System

Figure 2 presents the third level of granularity of the architecture of the DFR HLI DFClaaS system. The first level of architectural granularity is cloud computing architecture. The microservices architecture follows this at the second level. At the third level, the microservice architecture hosts layered architectural patterns. Finally, model-view-controller (MVC) architecture is presented at the fourth level of granularity.

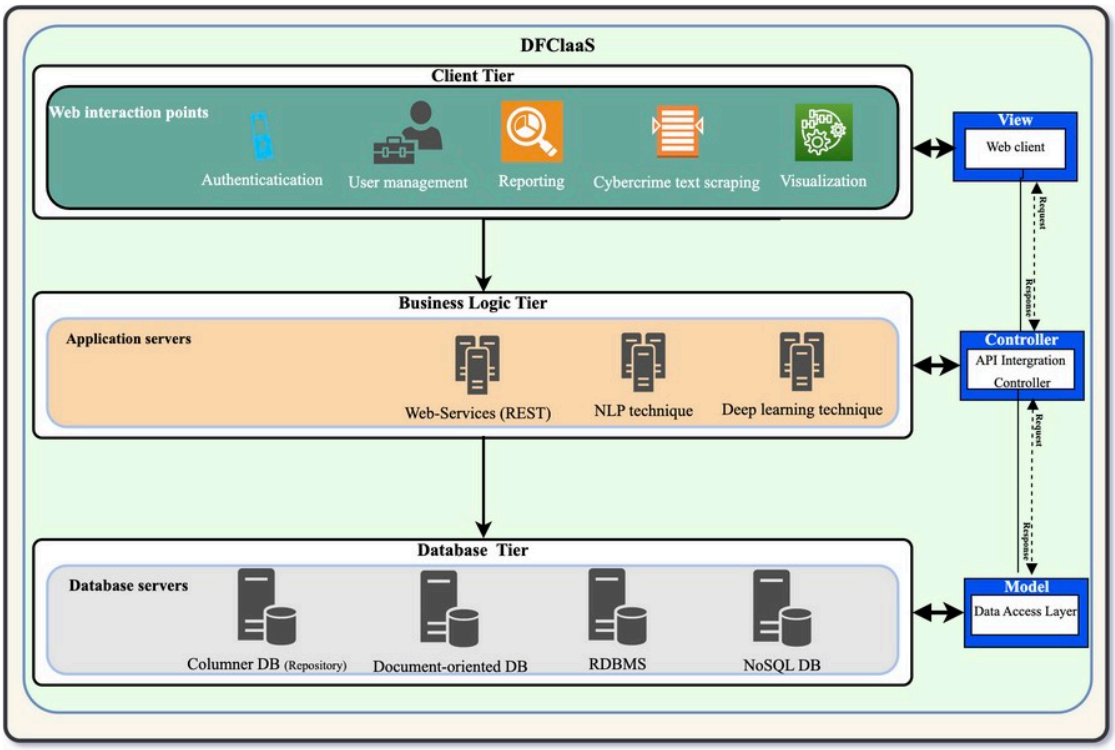


Figure 2: Layered Architectural Patterns of the DFR HLI DFClaaSystem

The architectures are displayed in Figure 2, showing layered architecture. The client tier consists of the actions that users can perform on the web interface of the DFR HLI DFClaaS system. It consists of the authentication, user management, cybercrime reporting, cybercrime text scraping and visualization. They are connected to the web client of the MVC architectural pattern (Adam et al., 2016). The MVC addresses the need for the front end of the DFR HLI DFClaaS system, which is designed for fast dynamic changes to ensure that data is displayed and visualized accurately. It provides a reactive functional design that encapsulates the system's core requirements to make the system usable and address the digital forensic readiness needs of the users. The view component implements the actual user interface to show the state of the model (see Figure 2), while also facilitating the user's interaction with other system components. The segregation of the DFR HLI system into a client, business logic, and persistence function ensures seamless service request to the server and the flexible reusable of functions. The business logic tier is the application servers layer, which consists of logic that collaborates to manage the system's working activities. They are web services, NLP and deep learning techniques. It implement digital forensic readiness and control security access. The API controller manages the functions that ensure the physical separation of the various layer while ensuring active communication interactions. The database tier hosts and manages the columnar database that maintains the cybercrime data repository, the document DB that handles any upload of cybercrime investigation documents by digital forensic investigation. The layered architecture further segregates the system into two applications, where the client makes a service request to the server.

The services interact in messaging, and the event-driven implicit invocation approach of message sourcing of events occurs in components of the microservices. The paradigm promotes the prevention, detection, consumption of, and reaction to each of the services' various activities and events when it encounters cybercrime text-data semantics previously flagged by the NLP components. The business logic layer contains the logic of the DFR HLI DFClaaS system, like the semantic, morphology and lexical businesses rule, gathering and identifying the logic within the system that determines how the behaviour correlates to other services. Finally, the database layer handles the storage, retrieval, and updating. The system database is designed to handle continually data

increase and are open-source and scalable to handle the large volumes of data, while continuous learning of the historical cybercrime data. As shown in Figure 2, the database tier (persistence) API allows interaction between the micro-services, the layered and the MVC architecture at the database level while maintaining a decoupled state. Next, the core quality requirements that must be addressed by the DFR HLI DFClaaS are discussed.

2.3.1 The Core Quality Requirements of the DFR HLI DFClaaS System

Quality requirements of any system are identified based on the functional requirements for the DFR HLI DFClaaS are shown in Table 1.

Security The security component focuses on the need to address confidentiality, integrity, and access control while complying with legal requirements of forensic soundness and chain of evidence rules. The security needs of the DFR HLI DFClaaS system depend on the end-to-end encryption of the communication channels to ensure that attacks, such as man-in-the-middle-attack are addressed. This ensures that every activity is traceable and retractable. Other aspects to be considered in the system security setup include: (i) The system should not read or decrypt any information in a message package that is not critical for delivering the message, cryptographic hash and digital signatures are some architectural strategies that addressed this (ii) Communications with the server are encrypted, critical messages are signed to prevent tampering, and messages between two users are encrypted. (iii) Regular monitoring and backups of the database are incorporated, and the SHA512 hashing algorithm is employed. In addition, peer-to-peer encryption of the communication channel, such as AES-256 is enabled at the coding level. (iv) Users authorization is determine their access levels and privileges. (v) Role-based access control ensures that users only authorized to interact with the system's aspects they have been granted permission. (vi) Unscrupulous users could flood the system with junk data, that could create denial-of-service attacks, cross-site request forgery, service side request forgery, and similar cyber-attacks.

Usability & Availability - It ensures that the application meets user expectations and improves the users' experience. It ensures intermittent, weak connections do not degrade the user's experience, maintains reliability and refers to the system's uptime. (i) The system is designed and developed always to provide real-time, updated information on the degree of availability of the system. (ii) in database failure or corruption, the server retrieves replacement and backups. (iii) The uptime and downtime of the system are recorded regularly to calculate and improve the system's availability. (iv) The server should only have a maximum of 0.0285% (2.5) or fewer hours of downtime in a year. (v) The overall system is designed to maintain at the minimum of 95% rating over 24 hours.

Reliability The attributes focus on the application's ability to continue working as it should, irrespective of possible latency. Therefore high fault tolerance is a built-in place to ensure that the system does not crash; thorough testing is to ensure durability at all pipeline and service levels. This is maintained by thoroughly implementing the following: (i) The database update process at the microservice level, and the layered architecture are be able to roll back updates when failure occurs. (ii) The application's SLA is estimated to be at 98% at its lowest operation level. If a crash happens, the system provide adequate data on why the crash occurred, allowing the administrator to fix the error, with no single point of failure with 99% uptime. The probability of a failure occurring on demand is less than 1 in 500 instances.

Performance The system is developed to be available at least 99% of the time, so the effect of fall-back, heavy-tailed traffic, and low latency are mitigated; therefore, a 95% up-time is expected. The application compensates with the response, process request and resource utilization, more specifically, the speed at which the system loads, this quality attribute is addressed to handle call requests and returns as follows: (i) The system should be able to process 1000 transactions per second. (ii) The system should have fewer than 3 seconds response time. This will be tested with performance testing. (iii) All web pages should be downloaded from the server within 5 seconds.

Architectural strategies involve making explicit trade-offs as to which feature of a quality requirement is more important to address the for the DFR HLI DFClaaS system.

Discussion

In the last few months after the COVID-19 pandemic, cybercrime incidents have risen, (Buil-Gil et al., 2021). This, therefore, means that individuals and businesses are the most affected. This study considered cybercrime in line with the victim's and the perpetrator's communication interaction. Irrespective of the type of cyber incident, be

it technical or non-technical, some form of string data communication occurred between the human users, providing insights to investigate the incident (Baror and Venter, 2019),(Baror et al., 2021). The criteria for evaluating cloud platforms and services are security, performance, scalability, adaptability, accessibility, and usability, which are the core quality attribute of the DFClaaS system is as follow:

- Allow users to report cybercrime either anonymously or otherwise. In the process generating valid historical cybercrime string to develop cybercrime semantics, lexicon and language using deep machine learning techniques.
- To learn and build knowledge features from the cybercrime string that show patterns consisting of identifiable cybercrime strings.
- The learned historical cybercrime string data detect and trigger units to notify users, subsystem or organizations of potential cybercrime.
- The distinguishing feature of the DFR HLI DFClaaS is the 'all-in-one' deployment. That is, the data analytical functions, the natural language processing techniques, deep learning and the digital forensic incident discovery services are integrated as one 'large' system. There is also the novel concept of anonymous cybercrime report, available historical string data stored and cybercrime language services.

This paper presents a functional software architectural design of a digital forensic readiness service based on human-language interaction. The DFR HLI DFClaaS system is especially advantageous for small and medium-sized enterprises (SMEs) during COVID-19, where most organizations opt to work from home to fulfil their customers' needs (Huaman et al., 2021).

The DFR HLI DFClaaS is designed using the URDAD method. This technology-independent architecture combines the system's complexity while presenting a high flexibility that allows for service and component re-usability. The DFR HLI DFClaaS is compartmentalized to address incident threat identification, incident threat trigger, incident threat isolation, incident threat investigation and the generation of cyber semantics for future cyber incident discovery. Therefore, developing the DRF HLI application with the overall architecture outlined in this paper is unique and essential for any human-centred digital forensic solution.

One drawback of the DFR HLI system is that user's must subscribe to DFClaaS service at the SaaS level of their public cloud usage to use the system. Similarly, the anonymous cyber-attacks reporting function is one challenge of the DFR HLI system as it is a potential loop hole to flood the system with junk string data and spike traffic. The DFR HLI DFClaaS system require a large volume of cybercrime data sets to increase accuracy. This is because deep learning techniques used for The DFR HLI DFClaaS system went a step further with a built-in level of security - i.e., authorization. The security architectural style requires users to satisfy the authorization requirements, which is a critical factor to verify users' actions and roles access services (Baror et al., 2020).

Conclusion

The digital forensic field tends to focus on the aftermath of an incident. Since the COVID-19 pandemic, small and medium-scale businesses (SMEs) have been affected (Parilla, 2021) by various cybercrime attacks.

The DFR HLI DFClaaS system is designed to gather cybercrime string data, learn the interaction behaviour of the cyber criminal's communications of the data and then provide a cyber semantics, lexicons and language preparedness using NLP and deep learning techniques. This is by measuring the sentimental, numeral, and contextual differences and anomalies of cybercrime data. Since the DFR HLI DFClaaS system is designed as a microservice-based architecture, each service is presented as an entity that is a callable API object. Therefore, integrating the DFR HLI DFClaaS system into any existing digital forensic application is easy.

References

Kai Adam, Arvid Butting, Robert Heim, Oliver Kautz, Bernhard Rumpe, and Andreas Wortmann. Model-driven separation of concerns for service robotics. In Proceedings of the International Workshop on Domain-Specific Modeling, pages 22–27, 2016.

Stacey O Baror, Hein S Venter, and Richard Adeyemi. A natural human language framework for digital forensic readiness in the public cloud. Australian Journal of Forensic Sciences, 53(5):566–591, 2021.

Stacey Omeleze Baror and Hein Venter. A taxonomy for cybercrime attack in the public cloud. In International Conference on Cyber Warfare and Security, pages 505–X. Academic Conferences International Limited, 2019.

Stacey Omeleze Baror, Richard Adeyemi Ikuesan, and Hein S Venter. A defined digital forensic criteria for cybercrime reporting. In International Conference on Cyber Warfare and Security, pages 617–XVIII. Academic Conferences International Limited, 2020.

David Buil-Gil, Fernando Miro-Llinares, Asier Moneva, Steven Kemp, and Nacho D´´ıaz-Castano. Cybercrime and shifts in ˜ opportunities during covid-19: preliminary analysis in the uk. European Societies, 23(sup1):S47–S59, 2021.

Yongzhen Chen, Yi Wang, Yajun Ha, Miguel Rodel Felipe, Shuqin Ren, and Khin Mi Mi Aung. saes: A high throughput and low latency secure cloud storage with pipelined dma based pcie interface. In 2013 International Conference on Field-Programmable Technology (FPT), pages 374–377. IEEE, 2013.

Julia Hirschberg and Christopher D Manning. Advances in natural language processing. Science, 349(6245):261–266, 2015.

Nicolas Huaman, Bennet von Skarczinski, Christian Stransky, Dominik Wermke, Yasemin Acar, Arne Dreißigacker, and Sascha Fahl. A {Large-Scale} interview study on information security in and attacks against small and medium-sized enterprises. In 30th USENIX Security Symposium (USENIX Security 21), pages 1235–1252, 2021.

Rick Kazman, Len Bass, Paul Clements. Software Architecture in Practice, 3rd Edition. Addison-Wesley Professional. Part of the SEI Series in Software Engineering series, 2012.

Stacey Omeleze and Hein S Venter. Testing the harmonised digital forensic investigation process model using an android mobile phone. In 2013 Information Security for South Africa, pages 1–8. IEEE, 2013.

Eric Santos Parilla. Effects of covid-19 pandemic on micro, small, and medium-sized enterprises in the province of ilocos norte philippines. In RSF Conference Series: Business, Management and Social Sciences, volume 1, pages 46–57, 2021.

Avinash Singh, Richard Adeyemi Ikuesan, and Hein Venter. Secure storage model for digital forensic readiness. IEEE Access, 2022.

Fritz Solms, Stefan Gruner, and Cuen Edwards. Urdad as a quality-driven analysis and design process. In New Trends in Software Methodologies, Tools and Techniques, pages 141–158. IOS Press, 2011.

Fritz Solms, Priscilla Naa Dedei Hammond, and Linda Marshall. Constraints-based urdad model verification. In ENASE, pages 148–155, 2016.

Aleksandar Valjarevic, Hein S Venter, and Melissa Ingles. Towards a prototype for guidance and implementation of a standardized digital forensic investigation process. In Information Security for South Africa (ISSA), 2014, pages 1–8. IEEE, 2014.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 93–104, Brussels, Belgium, 2019. Association for Computational Linguistics. doi: 10.18653/v1/d18-1009. URL https://www.aclweb.org/anthology/D18-1009.

Teaching Social Science Aspects of Information Security

Igor Bernik
Faculty of Criminal Justice and Security, University of Maribor, Slovenia

igor.bernik@um.si

Abstract: As information security has become increasingly crucial in our daily lives, there is a growing need to teach its social science aspects. This paper explores the challenges and best practices for teaching social science aspects of information security. It begins with the importance of information security and cyberspace and highlights the human aspects of information security. Next, it discusses the role of social science in understanding information security and how social science can help us better design and implement security measures. The paper identifies challenges in teaching social science aspects of information security, such as the interdisciplinary nature of the subject and the need for a standardised curriculum. Finally, the paper outlines best practices for teaching social science aspects of information security, such as using case studies and real-world examples, incorporating interactive and experiential learning, and leveraging existing resources. The conclusion highlights the importance of incorporating social science aspects of information security in education and suggests future research directions.

Keywords: Information Security, Cyberspace, Social Science, Human Aspects, Teaching

1. Introduction

In recent years, the importance of social science in information security education has become increasingly recognised. Information security is a complex field involving technical aspects and social and human factors (Moor, 2010; Singh et al, 2018). Several researchers have noted that human factors are crucial in information security, including how people interact with information systems (Kumar and Pang, 2015; Mitropoulos et al, 2018). Understanding the human factors contributing to security incidents is critical to developing effective solutions. Social science can help us better understand the human aspects of security, including the motivations and behaviours of attackers, the psychology of users and employees, and the cultural and organisational factors that influence security (Furnell and Warren, 2015; Mitropoulos et al, 2018). Human factors are influenced by the broader socio-technical context in which information systems operate (Clarke, 2014; Moor, 2010). This context includes organisational culture, social norms, and legal and regulatory frameworks. Therefore, effective information security requires technical solutions and consideration of these broader contextual factors.

At the Faculty of Criminal Justice and Security at the University of Maribor (UM FCJS), we have offered an undergraduate Information Security program since 2010. In addition to the essential technical aspects, our program focuses mainly on the sociological and criminological aspects of information security. Given the nature of the teaching of security studies at the time, we decided that our program would be different from most of the available programs and would focus not only on the technical but also on the security and social sciences aspects of information security, preparing students for the constantly changing technologies and the associated security issues in cyberspace. Therefore, we have provided the following definition of information security: "Information security is an interdisciplinary field that studies phenomena from the perspective of computer science, information systems, security studies, internal security and protection, develops and studies mechanisms that society uses to ensure security and privacy, and the protection of the state, ensuring security in the community, organisations, and individuals against behaviour that may harm the individual, community, or state, and develops theories and methods for protecting the information and privacy of the subjects as mentioned above." (UM FCJS, n.d.) The curriculum, in addition to the basic methodological and technical knowledge of information systems and information security, includes areas of (criminal, procedural, and administrative) law and government regulation, national security, management and administration, sociology, psychology, criminology, and criminality. Over the decade of implementation and changes to the program due to technological advancements and constantly evolving needs in the economy, we increasingly find that the sociological aspects of information security are equally important as the technical aspects. Based on our experience and literature studies, we provide some guidelines for preparing information security content and study programs that move away from traditional technical fields and towards security studies and management.

Social science can help us better understand the human aspects of security, including the motivations and behaviours of attackers, the psychology of users and employees, and the cultural and organisational factors that influence security (Gordon et al, 2017). With the research on vulnerabilities in information, systems were identified to understand how people interact with technology and develop more effective security policies and practices (Bada and Sasse, 2014). For example, social science research has shown that users often struggle to

understand complex security warnings and are likelier to click phishing links when distracted or under stress (Kumaraguru et al, 2013). This knowledge can inform the design of more user-friendly security interfaces and better security awareness training. Security decisions can significantly impact individuals, organisations, and society, and it is essential to consider these impacts. Social science research can help us to understand the trade-offs between security, privacy, and other values and to develop security policies that balance these competing interests (Van den Berghe and Walgrave, 2016). Including social science subjects in information security education can help students develop a more holistic and nuanced understanding of security and develop the critical thinking and problem-solving skills needed to address the complex security challenges of the future.

This paper explores the social science aspects of information security and provides educators with practices and strategies for teaching these aspects in their curriculum. It will also discuss educators' challenges in teaching social science aspects and offer suggestions for future research and improvement in information security education.

2. Understanding Social Science in Information Security

Social science aspects of information security refer to the broader socio-technical context in which information systems are situated, including the human and social factors that influence the security of information systems (Furnell and Warren, 2015). Factors include the human behaviours, attitudes, and motivations of individuals and groups, as well as the cultural and organisational context in which they operate:

1. Human behaviours: This factor refers to the actions that individuals take in their interactions with information systems, such as their adherence to security policies, their password management practices, and their use of personal devices for work purposes. Research (Kumar and Pang, 2015) has shown that human error is a major cause of information security incidents, making it crucial to understand and address human behaviours in information security.
2. Attitudes: Attitudes refer to individuals' beliefs and feelings about information security, including their perceived risks, trust in the system, and willingness to comply with security policies. Understanding individuals' attitudes can help organisations design effective security awareness training and communication strategies (Van Niekerk and Von Solms, 2017).
3. Motivations: Factor refers to why individuals may engage in risky behaviours, such as hacking or sharing sensitive information. Motivations can be influenced by financial gain, revenge, or a desire for recognition (Singer and Friedman, 2017). Understanding these motivations can help organisations design effective deterrence strategies and address underlying vulnerabilities.
4. Cultural context: Culture refers to the shared values, beliefs, and norms that shape individuals' behaviours and attitudes. Cultural factors can influence how individuals perceive and respond to security threats and shape organisational practices and policies (Singh et al, 2018). Understanding cultural factors is essential for designing effective security strategies in diverse settings.
5. Organisational context: This factor refers to the policies, procedures, and structures that shape how individuals interact with information systems within an organisation. Organisational context can influence the effectiveness of security controls and the willingness of employees to comply with security policies (Mangalaraj et al, 2016). Understanding the organisational context is essential for designing effective security policies and practices.

While technical solutions and tools are important in securing information systems, understanding the social and human factors that influence security is equally important. Social science research can help us better understand individuals' and groups' behaviours, motivations, and attitudes concerning information security and develop more effective policies and practices to improve safety and security.

3. Challenges in Teaching Social Science Aspects of Information Security

Teaching social science aspects of information security can be challenging for educators, as it requires them to address complex and multifaceted issues beyond purely technical concerns. Here are some of the challenges that educators may face:

1. Interdisciplinary nature: Social science aspects of information security draw on various disciplines, including security, psychology, sociology, and law. Educators may need a broad knowledge (Furnell and Warren, 2015; Singh et al, 2018) to teach social science aspects of information security effectively.

2. Lack of integration: Social science aspects of information security may not be well integrated into existing information security curricula. Educators may need to develop or modify new courses to address mentioned topics (Clarke, 2014; Mitropoulos et al, 2018).
3. Difficulties in measurement: Social science research often involves qualitative or subjective data, which can be more difficult to measure than quantitative data. This can make it challenging for educators to assess student learning and evaluate teaching methods' effectiveness (Furnell and Warren, 2015).
4. Limited resources: Educators may need more resources (Furnell and Warren, 2015; Mitropoulos et al, 2018), such as time, funding, or access to social science experts, to develop and implement effective teaching strategies.
5. Resistance to change: Educators may encounter resistance (Singh et al, 2018; Mitropoulos et al, 2018) from students and/or other stakeholders who are more focused on technical aspects of information security and may not see the value of social science approaches.
6. Rapidly evolving field: Social science aspects of information security are constantly changing as new technologies and threats emerge. Educators must stay up-to-date with the latest research and developments (Furnell and Warren, 2015; Clarke, 2014) to teach relevant and timely information.

Addressing these challenges requires educators to be creative, flexible, and collaborative. They may need to collaborate with social science experts and other educators to develop effective teaching strategies and to advocate for the importance of social science aspects of information security within their institutions and the broader community. Ways to overcome the mentioned challenges faced by educators:

1. Collaboration: Educators can collaborate with social science experts and other educators to develop interdisciplinary approaches to teaching social science aspects of information security. By working together, educators can leverage the strengths of each discipline and create more comprehensive and effective teaching strategies.
2. Integration: Social science aspects of information security can be integrated into existing information security curricula. This can involve modifying existing courses or developing new ones incorporating social science topics. Educators can also use case studies and examples that illustrate the relevance of social science to information security.
3. Active learning: Social science aspects of information security can be taught through active learning approaches that engage students in problem-solving and critical thinking. For example, educators can use group discussions, debates, and simulations that challenge students to analyse complex issues and make informed decisions.
4. Assessment: Educators can use various assessment methods to measure student learning, including written assignments, presentations, and practical exercises. Rubrics and clear learning objectives can help educators assess student learning outcomes and identify areas for improvement.
5. Professional development: Educators can stay up-to-date with the latest research and developments in social science aspects of information security by attending conferences, workshops, and webinars. Professional development opportunities can also help educators develop new teaching strategies and connect with peers in the field.
6. Advocacy: Educators can advocate for the importance of social science aspects of information security within their institutions and the broader community. This can involve engaging with decision-makers and stakeholders, highlighting the relevance of social science to information security, and showcasing examples of successful teaching strategies.

Above are strategies that educators can use to overcome these challenges and effectively teach the social science aspects of information security. Collaboration with social science and information security experts can help educators bring a deeper understanding of the social science aspects of information security into the classroom. Integrating social science topics into existing curricula can make the material more accessible and relevant to students. Active learning approaches, such as case studies and role-playing exercises, can engage students and help them apply social science concepts to real-world scenarios. Effective assessment methods can evaluate students' understanding of social science concepts and skills. Finally, ongoing professional development and advocacy for the importance of social science in information security can help educators stay current with the latest research and trends in the field.

Research in the field of education and information security supports these strategies. For example, a study by Maguire et al (2020) found that integrating social science topics into a cybersecurity curriculum increased students' interest in the subject and improved their critical thinking skills. Similarly, a study by Olson and Yousuf (2019) found that active learning approaches, such as problem-based learning and role-playing exercises,

improved students' understanding of social science concepts in information security. Teaching the social science aspects of information security is crucial for preparing students to address the complex challenges of the field.

4. Best Practices for Teaching Social Science Aspects of Information Security

Teaching information security requires a nuanced understanding of technical and social science aspects. However, educators need help effectively teaching the social science aspects of information security. To overcome the above challenges, there are presented best practices that educators can implement to effectively teach the social science aspects of information security. Effective teaching methods and strategies for teaching social science aspects of information security are:

1. Active learning: Active learning approaches, such as group discussions, problem-based learning, and case studies, effectively teach social science aspects of information security (Jøsang, 2017; Kim and Kim, 2019). These methods engage students in critical thinking and problem-solving and help them to apply social science concepts to real-world situations.
2. Multidisciplinary approaches: Social science aspects of information security require a multidisciplinary approach that draws on multiple disciplines, including psychology, sociology, anthropology, and law (Siponen, Vance and Willison, 2018). Educators can use multidisciplinary techniques to create more comprehensive and effective teaching strategies.
3. Experiential learning: Experiential learning approaches, such as internships, co-op programs, and service-learning projects, teach social science aspects of information security (Morrow and Brown, 2015). These methods provide students with hands-on experience and allow them to apply social science concepts in real-world settings.
4. Technology-enhanced learning: Technology-enhanced learning approaches, such as online courses, simulations, and virtual reality environments, effectively teach social science aspects of information security (Martin, 2017; Siponen et al, 2018). These methods provide students with flexible and interactive learning opportunities that can simulate real-world scenarios.

With the teaching methods and strategies, educators can effectively teach these topics and prepare students for a career in information security. With a deeper understanding of the social science aspects of information security, students can become well-rounded professionals equipped to address the complex challenges of the field. We can effectively integrate social science into the information security curriculum by employing methods and strategies. This integration provides students with a comprehensive understanding of the diverse and multifaceted nature of information security, enabling them to address the complex challenges of the field and contribute to the development of sustainable security solutions. Social science provides students with the tools and knowledge to analyse and manage these factors, including the psychology of cyber attackers, the impact of social norms and culture on security practices, and the ethical and legal implications of information security decisions. By integrating social science into the information security curriculum, students are better prepared to tackle the real-world challenges of the field and contribute to the development of effective and sustainable security solutions.

While there are various approaches to improving curricula, we recommend the following steps based on their proven success in practice:

1. Identify social science topics: The first step is identifying topics relevant to information security. These topics include human behaviour, ethics, psychology, organisational culture, and risk management.
2. Incorporate social science topics into existing courses: Social science topics can be integrated into existing information security courses, such as new technologies (artificial intelligence, Big Data etc.), network security, cybercrime and cyber warfare, crisis management and similar. For example, students could be asked to analyse the psychological motivations behind cybercriminal behaviour or to develop risk management strategies based on organisational culture.
3. Develop new courses: Institutions can also develop new courses that specifically focus on the modern aspects of information security. These courses could cover cybersecurity ethics, social engineering, and risk perception.
4. Partner with social science departments: Collaboration with social science departments can be beneficial in integrating social science into the information security curriculum. Joint courses or interdisciplinary projects can give students a more comprehensive understanding of information security.

At UM FCJS, we have successfully upgraded the curriculum by incorporating the presented teaching methods and strategies, allowing us to meet employers' demands. Our undergraduates' Information security program offers students a contemporary and exciting course of study that equips them with the knowledge necessary for success in their career development. As we move forward, we remain committed to further enhancing the program in line with the evolving trends and guidelines of the field. As a result, our graduates are highly sought after as experts and are recognised for their vital contribution to establishing contemporary business practices within their respective organisations.

5. Conclusion

The integration of social science aspects into information security education is crucial for producing well-rounded professionals who understand the human aspects of cybersecurity. By incorporating social science topics (Whitman and Mattord, 2016, Herath and Rao, 2019), such as human behaviour, ethics, psychology, organisational culture, and risk management, into the curriculum, educators can help students develop a more comprehensive understanding of information security. Educators need to work on teaching social science aspects of information security, such as the lack of interdisciplinary expertise, difficulty keeping up with rapidly evolving technologies, and students' reluctance to engage with social science topics (O'Connor and Lada, 2018). Several best practices can be used to overcome these challenges, such as active learning strategies, collaboration with social science departments, and real-world case studies (O'Connor and Lada, 2018). One of the examples is the introduction of database content as a technical element of information security, where data is also defined with a sociological and behavioural approach and understanding so that students understand the mathematical and technical connections between data and the impact of these on behaviour, perception and response of the individual dealing and working with data. Such an approach is geared towards enhancing students' understanding of this field.

Future research will focus on evaluating the effectiveness of these best practices in teaching social science aspects of information security, developing new interdisciplinary teaching methods, and exploring how emerging technologies such as artificial intelligence and blockchain can be integrated with social science topics. The focus would help create a more well-rounded and comprehensive curriculum that accounts for cybersecurity's social and technological aspects. Future research in these areas can further improve the quality of information security education and better prepare students for the complex security challenges of the future.

The importance of social science in information security education must be considered. By providing insights into the human factors that impact security and the broader societal implications of security decisions, social science can help to develop more effective security policies and practices. Educators can overcome the challenges of teaching social science aspects of information security by using effective teaching methods and strategies and integrating social science topics into the curriculum.

Acknowledgements

This publication is part of the Green and Resilient Transition for a Safe and Successful Society project, co-financed by the Ministry of Higher Education, Science, and Innovation of the Republic of Slovenia and the European Union's NextGenerationEU initiative.

References

Bada, M., and Sasse, M. A. (2014). *Research into usable security: Where are we now, and where do we go from here?* Proceedings of the 28th International BCS Human Computer Interaction Conference, pp 1-10.

Clarke, R. (2014). *Social issues and security. In Computer security, privacy and politics: Current issues, challenges and solutions* (pp 19-36). Springer International Publishing.

Furnell, S., and Warren, M. (2015). *Human factors in security. Cyber security: Management, systems, governance and assurance* (pp 127-146). Springer International Publishing.

Gordon, L. A., and Loeb, M. P. (2006). The economics of information security investment. ACM Transactions on Information and System Security (TISSEC), 9(4), pp 438-457.

Gordon, L. A., Loeb, M. P., Lucyshyn, W., and Zhou, L. (2017). A framework for using social science to improve cyber security. *Journal of Cybersecurity*, 3(1), pp 59-69.

Herath, T., and Rao, H. R. (2019). Social science in information security research: An interdisciplinary approach. *Information & Management*, 56(2), pp 225-235.

Jøsang, A. (2017). Teaching cybersecurity through gaming simulations. *IEEE Security & Privacy*, 15(3), pp 89-93.

Kim, Y., and Kim, H. (2019). Teaching cybersecurity ethics through case studies. *Journal of Computing Sciences in Colleges*, 34(2), pp 77-84.

Kshetri, N. (2017). Blockchain's roles in meeting key supply chain management objectives. International Journal of Information Management, 37(2), pp 1-7.

Kumar, S., and Pang, H. (2015). Exploring the human factors of information security incidents. *Computers & Security*, pp 50, 70-80.

Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., and Hong, J. (2013). Lessons learned from the world of unusable security interfaces. *IEEE Security & Privacy*, 11(1), pp 63-69.

Maguire, M., Donovan, E., and Chen, Y. (2020). Integrating social science topics into a cybersecurity curriculum. *Journal of Information Technology Education: Innovations in Practice*, 19, pp 273-287.

Mangalaraj, G., Nadarajan, R., and Dhuraisamy, R. (2016). A systematic review of organisational context factors influencing information security policy compliance. *Journal of Information Security and Applications*, 29, pp 50-72.

Martin, N. A. (2017). An exploration of virtual reality for teaching information security awareness. *Journal of Information Technology Education: Research*, 16, pp 207-227.

Meyer, E., and Canright, G. S. (2016). The flipped classroom model and problem-based learning: Improving patient safety through enhanced teamwork skills. *Journal of Professional Nursing*, 32(3), pp 191-197.

Mitropoulos, S., Dhillon, G., and Arachchilage, N. A. G. (2018). Human aspects of information security in organisations. *Journal of Business Research*, pp 82, 23-30.

Moor, J. H. (2010). Why we need better ethics for emerging technologies. *Ethics and Information Technology*, 12(1), pp 1-4.

Morrow, R., and Brown, J. (2015). Developing socially responsible information security professionals: Service-learning projects in the information systems security classroom. *Journal of Information Systems Education*, 26(3), pp 181-190.

O'Connor, L. T., and Lada, A. (2018). Active learning in cybersecurity education: A review. *Journal of Information Systems Education*, 29(4), pp 259-269.

Olson, J., and Yousuf, S. (2019). Active learning strategies for integrating social science into cybersecurity education. *Journal of Cybersecurity Education, Research and Practice*, 1(1), pp 19-32.

Singer, J. B., and Friedman, B. (2017). Motivations for hacking. *The Palgrave Handbook of Security, Risk and Intelligence* (pp 103-121). Palgrave Macmillan, London.

Singh, A. K., Bali, R. K., and Sharma, S. K. (2018). Human factors in information security: Review and future directions. *Journal of Organizational and End User Computing (JOEUC)*, 30(4), pp 1-20.

Siponen, M. T., Vance, A., and Willison, R. (2018). Educating employees about information security: Lessons from a survey of companies in Finland. Journal of Information Systems Education, 29(2), 85-96.

UM FCJS; (n.d.). *Undergraduate Study Programme in Information Security*. [online], Faculty of Criminal Justice and Security, University of Maribor, https://fvv.um.si/en/study/undergraduate-study-programmes/information-security/

Van den Berghe, L., and Walgrave, S. (2016). Public support for security policy: A review of the social psychological literature. *Journal of Police and Criminal Psychology*, 31(1), pp 15-28.

Van Niekerk, J. F., and Von Solms, R. (2017). A framework for security awareness training based on attitudes and motivation. *Computers & Security*, 70, pp 1-11.

Whitman, M. E., and Mattord, H. J. (2016). *Principles of information security*. Cengage Learning.

An Analysis of the MTI Crypto Investment Scam: User Case

Johnny Botha[1], Thor Pederson[2] and Louise Leenen[3]
[1]Council for Scientific and Industrial Research, Pretoria, South-Africa
[2]TCG Forensics, Cape Town, South-Africa
[3]University of Western Cape and CAIR, Cape Town, South-Africa

[1]jbotha1@csir.co.za
[2]thor@tcgforensics.co.za
[3]lleenen@uwc.ac.za

Abstract: Since the start of the Covid-19 pandemic, blockchain and cryptocurrency adoption has increased significantly. The adoption rate of blockchain-based technologies has surpassed the Internet adoption rate in the 90s and early 2000s. As this industry has grown significantly, so too has the instances of crypto scams. Numerous cryptocurrency scams exist to exploit users. The generally limited understanding of how cryptocurrencies operate has increased the possible number of scams, relying on people's misplaced sense of trust and desire for making money quickly and easily. As such, investment scams have also been growing in popularity. Mirror Trading International (MTI) has been named South Africa's biggest crypto scam in 2020, resulting in losses of $1.7 billion. It is also one of the largest reported international crypto investment scams. This paper focuses on a specific aspect of the MTI scam; an analysis on the fund movements on the blockchain from the perpetrators and members who benefited the most from the scam. The authors used various Open-Source Intelligence (OSINT) tools, alongside QLUE, as well as news articles and blockchain explorers. These tools and techniques are used to follow the money-trial on the blockchain, in search of possible mistakes made by the perpetrator. This could include instances where some personal information might have been leaked. With such disclosed personal information, OSINT tools and investigative techniques can be used to identify the criminals. Due to the CEO of MTI having been arrested, and the case currently being dealt with in the court of law in South Africa, this paper also presents investigative processes that could be followed. Thus, the focus of this paper is to follow the money and consequently propose a process for an investigator to investigate crypto crimes and scams on the blockchain. As the adoption of blockchain technologies continues to increase at unprecedented rates, it is imperative to produce investigative toolkits and use cases to help reduce time spent trying to catch bad actors within the generally anonymous realm of cryptocurrencies.

Keywords: Blockchain, Crypto scams, Crypto crime, Investigation, MTI, Process.

1. Introduction

Cryptocurrencies, Non-fungible tokens (NFTs), Decentralised Finance (DeFi) and smart contracts are all terms associated with blockchain technology. This technology has been growing at a rapid pace and provides unlimited applications. Some argue that Bitcoin gives power back to the people. However, the popularity of cryptocurrencies (crypto) has attracted the attention of many scammers and fraudsters. Undoubtedly, the rise of cryptocurrency has contributed to the immense increase in crime rates. Illicit transactions in cryptocurrency have reached a staggering $14-billion in 2021, an 80% increase from 2020 - this constitutes a new record (Sigalos, 2022) (Chainalysis, 2021). Scammers have always been around, but some of the characteristics of crypto are very appealing to them. Crypto has no middleman as in the case with banks. Instead, direct transactions occur between two individuals and transfers are much faster than traditional finance systems. When crypto is used on exchanges, the exchange can be seen as the middleman. Although many exchanges are still highly unregulated, they continue to act as the avenue to get crypto to fiat and vice versa. Pseudo names are being used on unregulated exchanges instead of actual personal details. It is difficult to trace crypto transactions and the space is relatively unregulated (Stylianou, 2022). Due to increasing regulations on crypto exchanges, some exchanges are gathering good *Know Your Customer* (KYC) information, others not at all (Lomas, 2023).

The Covid-19 pandemic brought difficult times with job losses and salary cuts. People became desperate to invest in alternative methods and crypto seemed like the perfect solution with the consequence that scammers took advantage of this opportunity (Xia, et al., 2020). Investigating and exploring cryptocurrency transactions remain intractably hard due to its pseudonymous nature and with every cryptocurrency having its own protocol and blockchain (Social Links, 2022).

Cryptocurrencies have become a popular target for criminals, and due to the nature of its design, criminals feel that they can commit crimes anonymously. However, cryptocurrencies are based on a public blockchain and visible to anyone. The flow of illicit transactions could be traced and investigated via advanced techniques with the goal of finding the destination cryptocurrency address that contains the stolen currencies. The next step as an investigator will be to unmask the owner of the address by combining Open-Source Intelligence (OSINT) and

KYC data collected from cryptocurrency exchanges. The investigator will then connect with law enforcement to have subpoenas issued in an attempt to seize or recover the stolen funds (Connors, 2022)

Some of the methods used by criminals to perform anonymous transactions on the blockchain are mixers or tumblers. These services combine cryptocurrencies of various users and send the crypto to another wallet. The investigator can then see that funds have been sent to the mixer and have been received by another person, but it breaks the direct connection between two individuals involved in the transaction. Criminals also make use of nested and unregulated exchanges that require no KYC and are not subject to Anti Money Laundering (AML) requirements. Criminals will often make use of privacy coins that are not tracable at all, making it almost impossible for investigators to trace the funds. Further more, Peer-to-Peer (P2P) decentralised crypto networks are being used, which also do not require KYC from customers. NFTs are often commonly used for money laundering. DeFi is the latest trend in the crypto space that has increased in popularity, and criminals have made use of this opportunity to send funds from illicit wallets (Stylianou, 2022).

Investors need to take all these methods into account when trying to connect the dots and follow the money. This paper analyses a specific, selected case on the Mirror Trading International (MTI) Platform. The case was selected from the website MTI-Leaks (MTI-Leaks, 2022). MTI-Leaks issued several publications with data on the founders of the scam as well as members who benefitted the most from the scam. In collaboration with TCG Forensic, a cyber-crime and digital forensics company based in South Africa, an initial process is proposed (section 3) on how certain crypto crimes can be investigated. This process is followed in section 4 where the case is analysed and investigated and concluded in section 5. Jeff Lomas, a detective and digital forensics examiner at the USA Las Vegas Metropolitan Police Department, kindly performed a review of the authors' investigation, and gives valuable insights on key focus points when dealing with criminal investigations (Lomas, 2023).

2. MTI Background

MTI was a network or multi-level marketing scheme that claimed to offer automated trading services via bots, on behalf of its members, in cryptocurrency derivatives. The scheme promised a consistent monthly 10% return to members. The website shut down in December 2020, and the CEO, Johann Steynberg, vanished and fled the country. Steynberg was arrested on 29 December 2021 in Brazil, and returned to South Africa, one year after his disappearance, for allegedly presenting fake identification to law enforcement officers. It was then discovered that he is a wanted person, the CEO of MTI, by South African Police Services (SAPS). At the time of writing, the legal process is still in progress. A trusted source with knowledge of the case has reported to MyBroadband that over 46, 000 Bitcoin passed through the scheme (Vermeulen, 2022).

TCG Forensics is one of the main investigators on several MTI cases. This paper has been written in collaboration with TCG Forensics and one of their main investigators on crypto crime cases, Thor Pederson. Section 3 explains the use case selection and gives a high-level overview of the MTI dashboard and platform functionality.

3. Use Case Selection

Anonymous ZA, a hacker group, released explosive details regarding the scheme, or rather the scam, on a Github link, MTI-Leaks (MTI-Leaks, 2022). Information such as their balance sheet, wallet addresses of founding members and members who benefitted the most from the scheme, the amount of Bitcoin allocated to the top addresses, withdrawals, cancelled and reversed withdrawals, remaining capital, the total money owed to members, etc. have been released. *Anonymous* also revealed that a handful of members received a "Founders Pool" bonus. Only 63 members were designated as founders, 0.038% of the total userbase (Vermeulen, 2020). This use case was selected from the MTI-Leaks website. The personal information of the target person will not be revealed but the initial address into which the target person invested will be revealed. It should be noted that the selected person of interest is not one of the founders of MTI, or the so-called masterminds of the scheme.

On the MTI Dashboard, the member was able see his/her portfolio with the total number of team members under the member's name, as well as the total direct referrals, the all-time bonusses, binary bonusses, the total wallet or account balance in MTI, and the total pool balance. The selected target person had 44419 team members, with 21 direct referrals, as illustrated in Figure 1. The member had 21783 members with 362.27 Bitcoin (BTC) in total on the left leg and 22636 members with 968.64 BTC on the right leg of the team (see Figure 2). Clearly the target person benefitted immensely from the scheme and qualifies as a person of interest. A high-level investigation has been conducted to trace the funds that were added/withdrawn to/and from the scheme.

Figure 1. Total Team Members and Direct Referrals

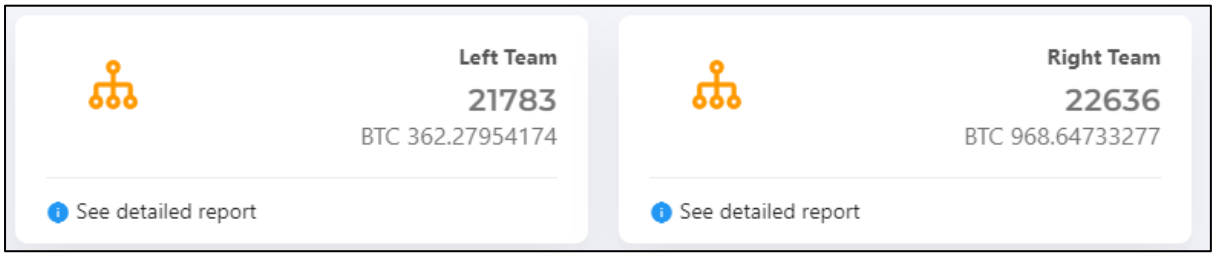


Figure 2. Left and Right Leg of Team

From the MTI dashboard, the user had a wallet section where an audit of funds added and withdrawn was kept. The investigation therefore has two angles:

1. Tracing the funds starting from the add funds address(s), the **Inputs**, and
2. starting from the withdrawal addresses(s), the **Outputs**.

The "**Add Fund History**" menu option on the dashboard (Figure 3) shows all five records recorded on the platform were selected as **inputs**. Note that the wallet address and transaction ids (TXIDs) are from the Bitcoin blockchain (

Table 1).

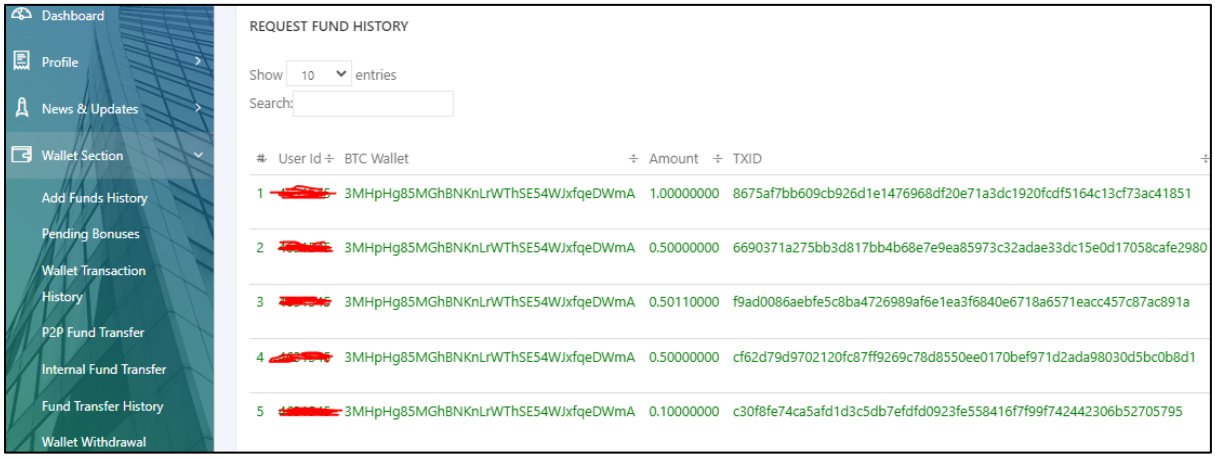


Figure 3. MTI Dashboard Menu (From MTI-Leaks)

Following the "Wallet Withdrawal" menu option from the dashboard, the very first withdrawal record was selected, Out-Address-1. However, this address had no TXID linked to it on the blockchain. A second address was selected, the first one from the list with a TXID linked to Out-Address-2. The third address selected was the most address used for pay-outs (Out-Address-3), with five transactions to start off with tracing the funds (see Table 2). These addresses are placed within the Outputs angle of the investigation.

According to Pederson, from TCG Forensics, the process for investigating a cryptocurrency scheme is dependent on the information available both publicly and during the course of the investigation. Before beginning the investigation, it is of utmost importance to collect all information that will enable the investigator to develop, compile and expose a timeline. An example of this could be a specific user or investor in a scheme that may have information very pertinent to the investigation process. An investor knows that over the course of the operation of the scam, the scam required deposits into a specific cryptocurrency address and that address changed over the course of a few months. This information alone information is extremely valuable and has a massive impact on where and when to investigate, and to find or join links over time (TCG Forensics, 2022).

Table 1. Inputs

Wallet Address	TXID	BTC	Date
3MHpHg85MGhBNKnLrWThSE54WJxfqeDWmA **(In-Address-1)**	8675af7bb609cb926d1e1476968df20e71a3dc1920fcdf5164c13cf73ac41851 **(Input-TXID-1)**	1	2020.05.17
3MHpHg85MGhBNKnLrWThSE54WJxfqeDWmA **(In-Address-1)**	6690371a275bb3d817bb4b68e7e9ea85973c32adae33dc15e0d17058cafe2980 **(Input-TXID-2)**	0.5	2020.03.29
3MHpHg85MGhBNKnLrWThSE54WJxfqeDWmA **(In-Address-1)**	f9ad0086aebfe5c8ba4726989af6e1ea3f6840e6718a6571eacc457c87ac891a **(Input-TXID -3)**	0.5011	2020.03.29
3MHpHg85MGhBNKnLrWThSE54WJxfqeDWmA **(In-Address-1)**	cf62d79d9702120fc87ff9269c78d8550ee0170bef971d2ada98030d5bc0b8d1 **(Input-TXID -4)**	0.5	2020.03.21
3MHpHg85MGhBNKnLrWThSE54WJxfqeDWmA **(In-Address-1)**	c30f8fe74ca5afd1d3c5db7efdfd0923fe558416f7f99f742442306b52705795 **(Input-TXID -5)**	0.1	2020.03.21

Table 2. Outputs

Wallet Address	TXID	Amount (BTC)	Date
1KX5f1mz8MBuCCdUyDcyRiZLH53v8cJjav (First address – **Out-Address-1**)	Not available	1.005	2019.12.12 23:06:20
3CvyH1syeBrJWLXkXpXtQtJ82wVq5naocV (**Out-Address-2)** (First Address with TXID)	7572ccc5c31b65e2eb72f62d78d42ec3ca20cf93884ec0e7664722ed54c4aa3c **(Output-TXID -1)**	2.5	2020.05.11 21:43:39
17AVMgsdxb7tELKwjn4fK3MymGkUTmY6DB (**Out-Address-3)** Most common used address	58c0b545a0eac027d143eb0838fe3b59fb093d1d383124b2f80893aaf00e697d **(Output-TXID -2)**	0.3	2020.07.19 22:58:01
	734e6394a4bda9de8707cdcca2e8760ef028f4a33b6845b32923ed918170a91c **(Output-TXID -3)**	0.27867493	0.27867493
	b92242994510d45a2dd12c1f9f2f89034f1870252839d066f6050db951153d12 **(Output-TXID -4)**	2020-07-20 03:00:25	2020.07.20 03:00:25
	d3d743092c6968cc255c631f0179b000e0a664328642e8921ace3c6ca30e95e2 **(Output-TXID -5)**	0.61480277	2020.07.26 22:47:06
	885700b5c8978d1c0267244a9a69ab41009e2ef1569a611c4a8d535321d35eec **(Output-TXID -6)**	0.1	2020.07.28 21:17:17

The second most important element is the flow of funds. The flow of funds is defined as the start and end points. In simple terms this means where the funds started, where they travelled, what happened to them and where they exited and all common points for such funds along the journey. Bad actors will deliberately attempt to obscure the flow of funds from investors and interested parties. The flow of funds is dependent on the type of scheme that was in operation. In the case of an investment scheme, one would expect to follow the path of the deposited funds to an investment vehicle on the blockchain, and this in itself may take on many forms such as crypto mining, token exchanges, trades, etc. Furthermore, it is not always possible to extrapolate the truth without confirmation that what one may find on the blockchain will concur with what was advertised to the investors (TCG Forensics, 2022).

4. High-level Proposed Process

Several engagements took place with TCG Forensics to determine a high-level process an investigator could follow to investigate a crypto scam. The focus was specifically on the MTI scam with the given data obtained from MTI-Leaks, but it could be applied to various forms of crypto scams, crimes, money laundering and fraud.

Seven steps have been identified as an initial high-level process, applicable for both the input and output addresses. The steps are outlined listed below and outlined in**Error! Reference source not found.**:

1. Identify the input addresses and the output or pay-out addresses.
2. Follow the funds on the blockchain starting from the identified addresses.
3. Identify key or common addresses used, their function and determine if they are linked to exchanges.
4. Further follow the funds from the key or common addresses identified.
5. Determine if there are any correlations between the addresses identified, such as a certain address being used as a destination address, linkages to exchanges, etc.
6. Identify the destination addresses.
7. At this point, legal support is required. Below are five key points to address when drafting legal charges:
    a. Draft a forensic report detailing the scam.
    b. Report to the Financial Sector Conduct Authority (FSCA).
    c. Draft the legal charges.
    d. Formulate and issue a subpoena on the identified linked exchanges and obtain the KYC identification documents.
    e. Freeze the funds linked to the destination addresses if possible.

Figure 4 illustrates the process.

5. Analysis and Investigation of the Selected MTI Case

The high-level process detailed in section 1, and drafted in **Error! Reference source not found.**, was used as a base and a guide to perform the analysis and investigation on the selected target person, who was a member of the MTI Ponzi scheme and benefitted tremendously out of the scheme. The target person is expected to repay the funds after the investigation and the required legal steps have been taken to finalise this case.

The crypto investigation tool, QLUE, was used to perform the investigation and to follow or trace the funds on the blockchain. (QLUE, 2022) Inputs and outputs almost never come from the same source (TCG Forensics, 2022). It is very difficult without sufficient context to know to whom a Bitcoin address belongs when using a crypto wallet (Lomas, 2023). However, the aim of this investigation is to follow the transaction flow on the blockchain, to determine patterns, to find key addresses where money was transferred to and from, and identify when a transaction was made to an exchange.

5.1 Inputs

The first part of the investigation is to investigate the funds being added onto the platform and by following the input transactions up to a point where they are linked to an exchange and a person can be identified.

Step 1

The input addresses have been identified from the MTI "**Add Fund History**" option and are listed in Table 1. Also refer to Figure 3.

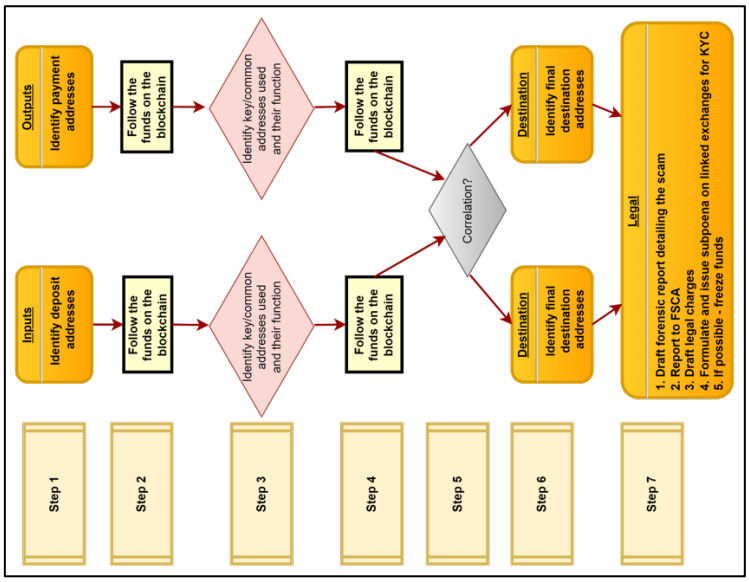


Figure 4: Investigation Process

Step 2 & Step 3

This investigation starts off with entering the input address, **In-Address-1**, into QLUE. Upon entering the input address, **In-Address-1**, QLUE returned no results and this immediately raised a red flag. The absence of results can mean no such address exists on the Bitcoin blockchain or no transaction has been made into the address. The next step was to confirm this on the online blockchain explorer, Blockchain.com (Blockchain.com, 2022). When entering the address into the explorer, the address does exist, but no transaction was ever made to this address. This may indicate the MTI administrators possibly intended to use this address as a dummy, and never reveal to the public where the funds actually went.

When entering the first input TXID, **Input-TXID-1** from
Table 1, it was discovered that the funds went to a different address than what was indicated on the MTI platform, and that it was a Bitcoin address that had been flagged by QLUE as a MTI address (refer to Figure 5 below - note that the green arrow indicates funds being received and the red arrow indicates funds being sent). This address will be referred to as **MTI-1**, as indicated in Figure 5. When following the funds further, another interesting discovery was that the funds went out of **MTI-1** to another address, on the very same day, approximately 3.5 hours later. Hereafter, the funds were sent to thousands of addresses. This indicates that mixer or tumbler services were possibly used.

When analysing **Input-TXID-2** and tracing the funds, it was revealed that the funds were transferred to a different MTI Bitcoin address, **MTI-2**, from **Input-TXID-1**. **Input-TXID-3** also went to **MTI-2** (refer to Figure 5).

Input-TXID-4 went to another MTI address, **MTI-3**, and **Input-TXID-5** went to **MTI-4** (refer to Figure 6).

Step 4 & Step 5

Upon further investigation and tracing of the funds, the tool revealed that thousands of transactions occurred after funds had been moved to these identified addresses. Considering **MTI-1** as an example, it indicates that from this address alone, 6K more transactions occurred. From **MTI-2**, 2K additional transactions occurred. Upon expanding **MTI-2**, only 53 transactions were made visible in QLUE, due to filtering applied, out of the 2K transactions shown in Figure 7. This indicates that MTI was possibly making use of crypto mixing to anonymise the crypto transactions. Crypto mixing or crypto tumblers is the process of obscuring the origins of a crypto transaction through mixing one's tokens with others of the same type. This results in thousands of transactions, with the goal to confuse the investigator, leaving multiple trails and paths of transactions on the blockchain (CryptoHopper, 2022). By using the tool QLUE, these traces have been followed and further key or common addresses have been identified. Key findings from the inputs were that funds were sent to a few common addresses linked to the MTI scam.

Step 6

Another noticeable finding was that after thousands of transactions had been made to several MTI linked addresses and other random addresses, funds were transferred to South-African exchanges such as Luno, VALR and AltCoinTrader. Funds were also transferred to exchanges outside of SA, such as Binance, Coins, Nexo, etc. The addresses on the South African exchanges have been marked in this investigation as the destination addresses, **Destination1** and **Destination2** (refer to Figure 7).

Step 7

Depending on the jurisdiction there are legal steps one is obliged to follow which are common to most countries. As such inter-agency or inter country co-operation can and often is requested by the investigator normally based on the submission of a thorough blockchain forensics report and backed by law enforcement authorities. Depending on the country from which cooperation is requested, a specific process may be required to be followed. Typically, the process may involve issuing subpoenas on the relevant exchanges via their compliance officer which often results in the investigator obtaining the required personal information from the identified exchanges. Once the target person has been identified, a notice should be sent to the person of interest, with a warning and court order that the funds linked to the person's name in certain specific identified addresses are not allowed to be moved. An instruction could be given to the exchanges to freeze the account for a time period, this normally requires the submission of the correct court processed documentation and proof of legal authority to act on behalf of the aggrieved. Should the account holder reside in the country where the investigation is being conducted, a court application can be made in order to instruct the account holder to cease and desist

from all activity on the blockchain account pending a court action against them, and violation of the court order is often backed up by contempt of court charges. In this event, the account holder will get a legal notice not to move any of the funds within these addresses (TCG Forensics, 2022). It should be noted that these people could highly likely be low-level mules and located in a different jurisdiction. To make an arrest, more context and other thresholds will be needed to get to the actual arrest point (Lomas, 2023). From this point, it becomes a legal case. If a decision is made by the court, funds could be moved and recovered as instructed by officials.

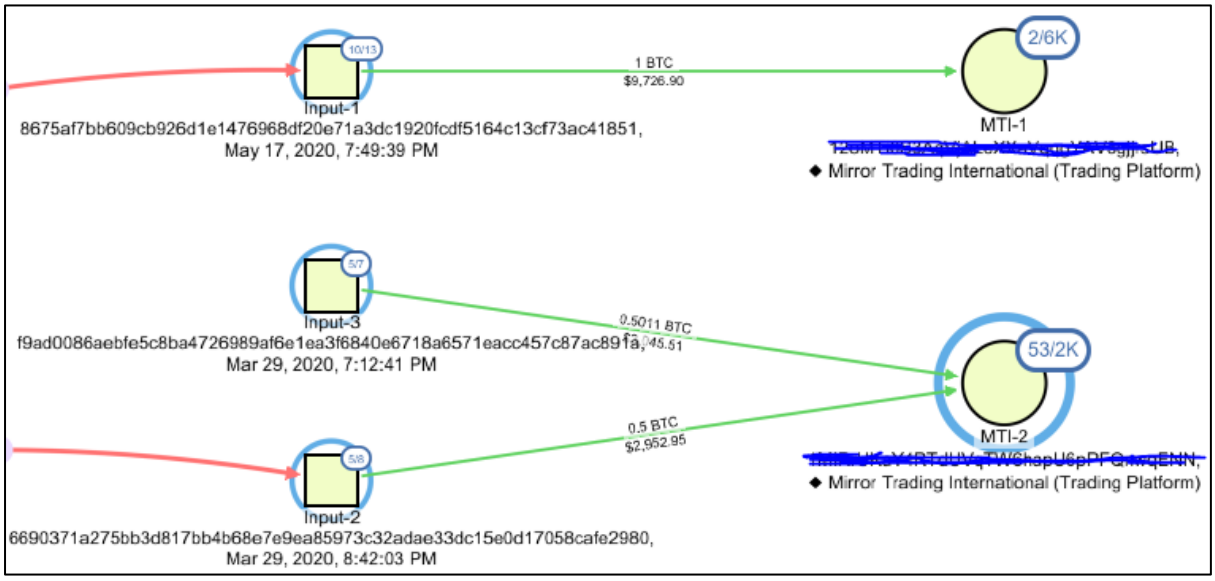


Figure 5. Input-1, 2 & 3

5.2 Outputs

The second part of the investigation is to investigate the pay-outs or outputs by following the output transactions up to a point where a person can be identified.

Step 1

The output addresses have been identified from the MTI Dashboard "Wallet Withdrawal" menu option and listed in Table 2. Also refer to Figure 3.

Step 2 & Step 3

By entering the first output address, Out-Address-1, into QLUE and following or tracing the funds, the tool revealed that four transactions are linked to this address even though MTI indicated no transaction is linked to this address; two input transactions with 0.005BTC and two output transactions with the same amount were found. MTI indicated that the amount that was transferred into this address is 1.005BTC. The sum of the four transactions found on the blockchain does not add up to this amount from the MTI platform, which already raises some questions. The two input transactions to Out-Address-1 that were identified are:

- b169639976a598029999191e0fc255114169286da756ce421f31e92620fe24b5 (Out-Address-1-Input-TXID-1) and
- 372f1b3f6b8d5c4f504a0cdb6371c427464e119d4382eb70adefbdecdaf6c120 (Out-Address-1-Input-TXID-2).

These two transactions would need to be backwards traced to determine from where the funds came. Upon tracing Out-Address-1-Input-TXID-1 for several transactions backwards, it was discovered there are several links to local South African exchanges as well as international exchanges. Upon tracing Out-Address-1-Input-TXID-2 backwards, no destination addresses could be identified, however, a single path could be followed that was ultimately linked to several exchanges.

The two output transactions to **Out-Address-1** that were identified are:

- dc4de50957a2a64a84a2a7d410155a6456083d365696fed9db0af6f015f62b40 (Out-Address-1-Output-TXID-1) and

- c8c4c08467e86fbe572d130c268b94711d51f475c196a76b32d949c3cb8ce5d8 (Out-Address-1-Output-TXID-2).

These two transactions need to be traced forwards in order to determine where the funds moved to, with the aim of finding a destination address.

When moving on and tracing **Output-TXID-1**, QLUE revealed that the funds were moved to blockchain addresses on both the Luno and AltcoinTrader platforms (see Figure 8). However, none of the addresses clustered in Luno and AltcoinTrader matched the address shown on the MTI dashboard that is supposed to be linked to **Output-TXID-1**. **Out-Address-2** is in fact linked to the Bitmex exchange and not linked to **Output-TXID-1** at all.

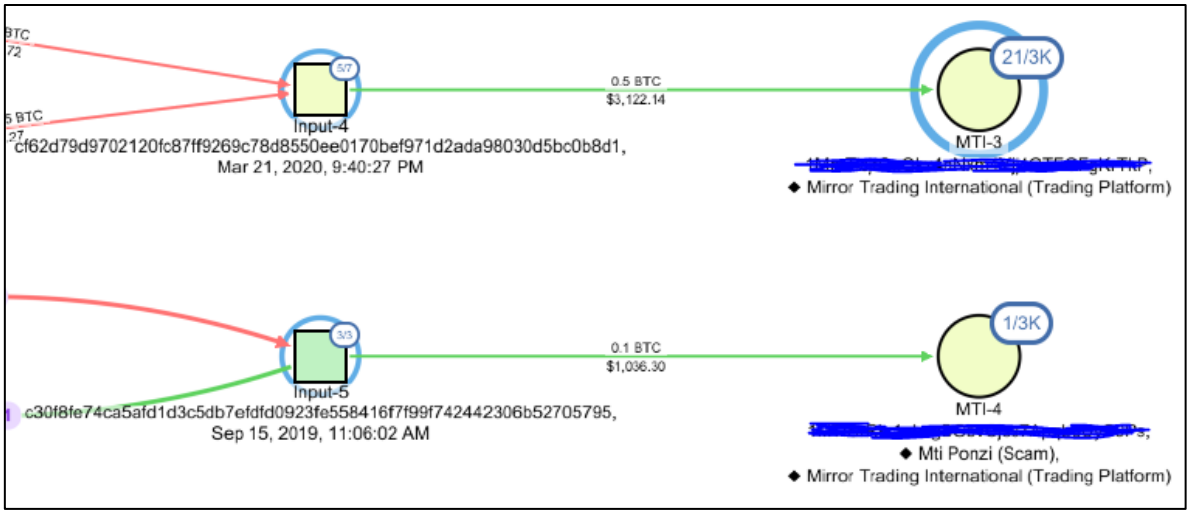


Figure 6. Input-4 & 5

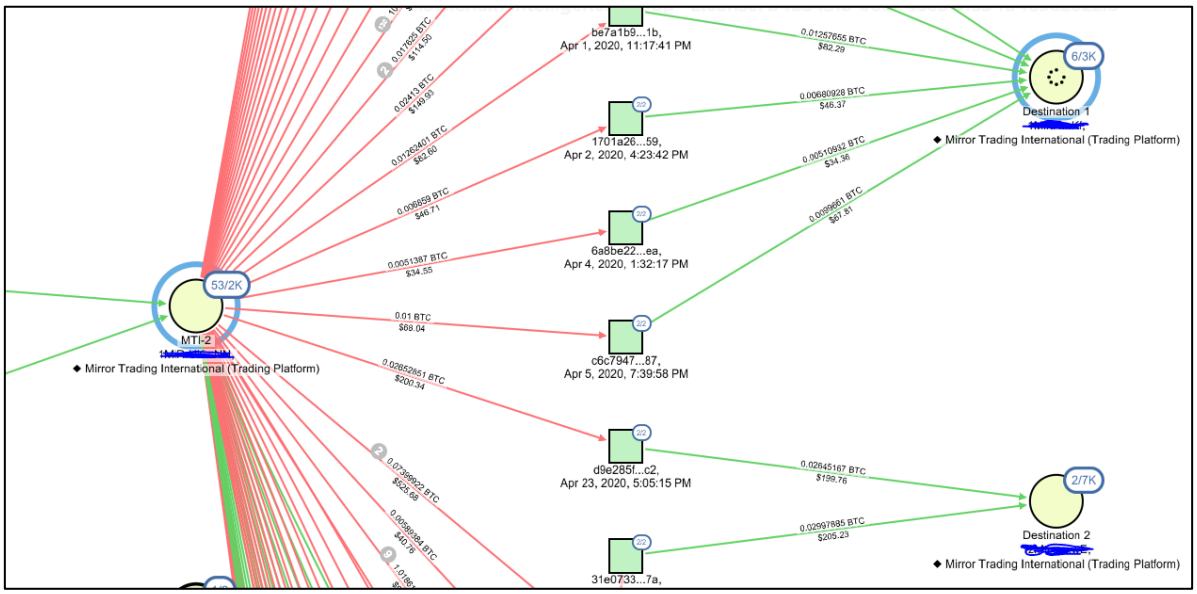


Figure 7. MTI-2 -> Destination 1 & 2

When tracing the most common used address from the MTI platform on the target person's profile, **Out-Address-3**, it revealed that the address was indeed linked to transaction **Output-TXID-5**, but not linked to any of the other transactions that MTI indicated are linked to the address. This indicates that the information shown to the members on the MTI dashboard is incorrect in this instance. The address has also been flagged by QLUE as a known MTI user address.

When tracing Output-TXID-2 via QLUE, no results came up. Another test was done using the tool Maltego (Maltego, 2023) with the Tatum transform for integrating with the blockchain. By tracing the input and output addresses via Maltego, several addresses were linked to the transaction (Figure 9), however none were the address shown on the MTI dashboard (see

Table 2). These addresses have been flagged for further investigation.

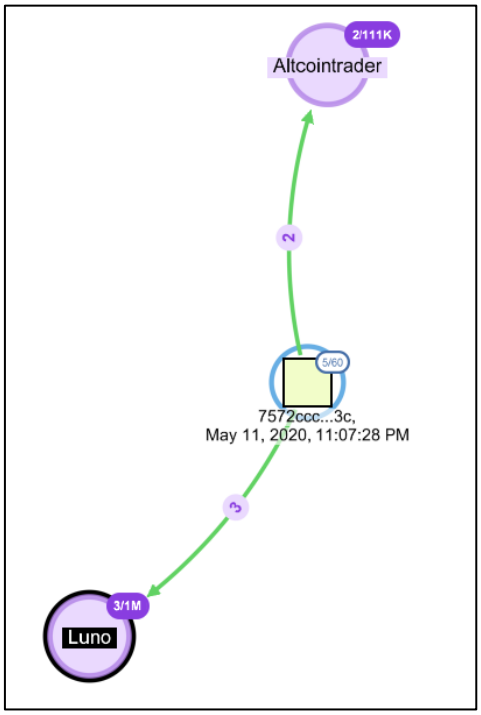


Figure 8. Output TXID-1

Step 4, Step 5 & Step 6

After further tracing on the **Out-Address-1-Input-TXID-1** path, a few addresses have been identified and linked to a number of exchanges such as Binance, Coinhako, Altcoin Trader, Unocoin wallet exchange and Coins exchange. When tracing the path backwards from **Out-Address-1-Input-TXID-1**, a correlation could be made to the exchanges Coinpayments, Bitmex, Bitpoint and Binance, as well as links to Dark Market activities based on the flagged addresses by QLUE. These addresses have been flagged as destination addresses for further investigation.

By following **Out-Address-1-Output-TXID-1** forward, several links to exchanges such as Bitstamp, BitPay, Paymium, Bitrex, Bitmex, Binance, Changelly, Hitbtc, Bitcoincom and Bithoven have been discovered. When tracing **Out-Address-1-Output-TXID-2** forward it was discovered that the two output transactions have a correlation on a particular transaction id that serves as an output for **Out-Address-1-Output-TXID-2** and is also an input for **Out-Address-1-Output-TXID-1**. QLUE has also clustered the two addresses together from the two outputs, which could indicate that these addresses belong to the same exchange or are controlled by the same person. These addresses have been flagged as destination addresses for further investigation.

The Bitmex address, **Out-Address-2** has been flagged as a destination address. The addresses linked to Luno and AltcoinTrader from **Output-TXID-1** have also been marked as destination addresses for further investigation.

The addresses linked to **Output-TXID-3**, **Output-TXID-5**, **Output-TXID-5** and **Output-TXID-6** have all been marked as destination addresses for further investigation. The addresses linked to **Output-TXID-6** have been flagged with a high priority as they indicate links to a known MTI user address as well as links to victims of fraud.

Step 7

Depending on the legal jurisdiction there are legal steps one is obliged to follow or request inter-agency co-operation - refer to section **Error! Reference source not found.**, Step 7 (TCG Forensics, 2022). By issuing subpoenas to the relevant exchanges via the compliance officer, the investigator could obtain personal information from the identified exchanges. If two or more of the addresses have the same personal information, there is a high possibility that the personal information is that of the target person. Similar to Step 7 in section **Error! Reference source not found.**, an instruction could be given to the exchanges to freeze the account. Legal notices and/or charges will follow, and the court will come to a final decision on what should happen to the funds. Again, it should be noted that these identified target persons could possibly be low-level mules and not the main target person of the scam (Lomas, 2023).

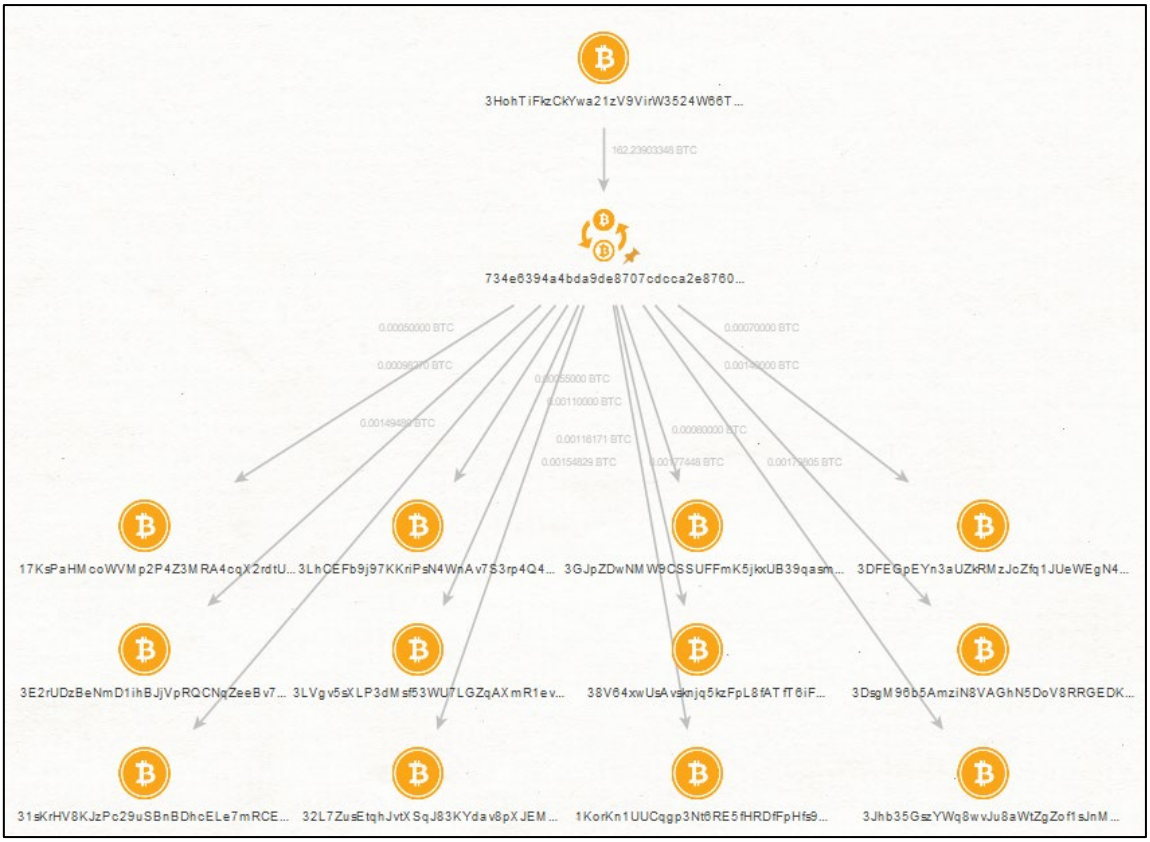


Figure 9. Output-TXID -2 - Linked Addresses

6. Conclusion

This paper focusses on a selected target person that was a member of the MTI scam who has benefitted immensely from the scam. MTI-Leaks has been used for the case selection and the data on input and output transactions and the addresses linked to them. Several discussions took place between the first author and a of crypto crime investigator from the South African company, TCG Forensics. The aim of the discussions is to propose a high-level process on how to investigate crypto scams and crimes. This process is presented in this paper and used as a base line to perform a high-level investigation. Limited free access to the tool QLUE has been obtained to perform the investigation. In addition, Jeff Lomaz, an OSINT and crypto crimes investigator from the USA, gave valuable insights on online criminal investigations.

The investigation followed two approaches, one for the funds being added onto the MTI platform, referred to as the inputs, and the second approach considers the fund withdrawals from the platform, referred to as the outputs, which are the pay-outs. After numerous efforts to trace the funds on all Bitcoin addresses and transactions that were selected from the MTI dashboard on the target person's profile, several addresses have been marked as the destination addresses and for further investigation. After the identification of the destination addresses, depending on the jurisdiction, subpoenas could be issued to the linked exchanges in order to obtain personal information on the individuals who performed the transactions on the particular exchanges. Further legal steps would then take place until a decision has been made by the court.

The crypto industry continues to grow, and criminals will continue to scam individuals and launder proceeds of their crimes using cryptocurrencies. The continued advancements in the blockchain and crypto industry make it more challenging for law enforcement agencies to detect and fight criminal activity. At the same time, it will be even more challenging for criminals to launder illegally obtained funds. It is important for governments, competent authorities, and law enforcement agencies to stay up to date with the methods adopted by criminals and to take advantage of blockchain analytic tools to effectively combat money laundering with the use of cryptocurrencies and crypto assets.

References

Blockchain.com. (2022, 12 14). *Blockchain.com - Explorer*. Retrieved from Blockchain.com: https://www.blockchain.com/explorer

Chainalysis. (2021). *The 2021 Crypto Crime Report.* Chainalysis. Retrieved from https://go.chainalysis.com/rs/503-FAP-074/images/Chainalysis-Crypto-Crime-2021.pdf

Connors, C. (2022, 12 09). *What is a Cryptocurrency Investigation?* Retrieved from ERMProtect.com: https://ermprotect.com/blog/what-is-a-cryptocurrency-investigation

CryptoHopper. (2022, 8 23). *What is crypto mixing?* Retrieved from www.cryptohopper.com: https://www.cryptohopper.com/news/7547-what-is-crypto-mixing

Lomas, J. (2023, 01 12). *Certified Instructor - Jeff Lomas*. Retrieved from sans.org: https://www.sans.org/profiles/jeffrey-lomas/

Maltego. (2023, 01 14). *Home page*. Retrieved from maltego.com: https://www.maltego.com/

MTI-Leaks. (2022, 11 24). *Mirror Trading International Data Leak*. Retrieved from MTI-Leaks: https://mtileaks.github.io/

QLUE. (2022, 11 25). *QLUE*. Retrieved from qlue.io: https://qlue.io

Sigalos, M. (2022, 01 7). *Crypto scammers took a record $14 billion in 2021*. Retrieved from cnbc.com: https://www.cnbc.com/2022/01/06/crypto-scammers-took-a-record-14-billion-in-2021-chainalysis.html

Social Links. (2022, April 22). *Enhancing Cryptocurrency Investigations with OSINT*. Retrieved from https://blog.sociallinks.io/: https://blog.sociallinks.io/cryptocurrency-investigations/

Stylianou, A. (2022, 4 4). *CRYPTOCURRENCIES - THE CHALLENGES FOR CRIMINALS AND INVESTIGATORS*. Retrieved from www.gci-ccm.org/: https://www.gci-ccm.org/insight/2022/04/cryptocurrencies-challenges-criminals-and-investigators

TCG Forensics. (2022, 11 28). *Home page*. Retrieved from www.tcgforensics.co.za: https://www.tcgforensics.co.za/

Vermeulen. (2022, 4 12). *Good news for Mirror Trading International scam victims*. Retrieved from mybroadband.co.za: https://mybroadband.co.za/news/cryptocurrency/440928-good-news-for-mirror-trading-international-scam-victims.html

Vermeulen, J. (2020, 9 20). *Explosive information about Mirror Trading International released by Anonymous ZA*. Retrieved from https://mybroadband.co.za/: https://mybroadband.co.za/news/business/368165-explosive-information-about-mirror-trading-international-released-by-anonymous-za.html

Xia, P., Wang, H., Luo, X., Wu, L., Zhou, Y., Bai, G., . . . Liu, X. (2020). Don't Fish in Troubled Waters! Characterizing Coronavirus-themed Cryptocurrency Scams. *2020 APWG Symposium on Electronic Crime Research (eCrime)*, (pp. 1-14). doi:10.1109/eCrime51433.2020.9493255

How to Safely Communicate with a Phishing Attacker by Email?

Ladislav Burita, Aneta Coufalikova and Kamil Halouzka
The Department of Informatics and Cyber Operations; Faculty of Military Technology, University of Defence; Brno, Czech Republic

ladislav.burita@unob.cz
aneta.coufalikova@unob.cz
kamil.halouzka@unob.cz

Abstract: The published study is a part of the long-term research of emails with phishing attacks against the article's author. In the previous three years, 3 experiments were carried out to analyze phishing emails. The result is their detailed classification. The subsequent experiment was focused on defense against phishing attacks using the rules of the MS Outlook email client. The last experiment, which is the article's content, is devoted to analyzing communications with phishing attackers. A fake identity was created for the experiment and security rules were set up. A total of 100 phishing emails were answered, with a preference for those whose content was not aimed at fulfilling any request; that was clarified during the communications. The conducted literature search confirmed the assumption that no one is engaged in similar research, so the results of the research may be more interesting for the cybersecurity community. The articles of the literary research are focused on the issue of social engineering from an interdisciplinary perspective. A great deal of attention has also been oriented on the influence of social networks on people information perception or on their exploitation in cyber-attacks. The result of the study is a statistical analysis of communications and a detailed analysis of its content. Out of 100 replies to the phishing email, 32 (32%) were answered by the phisher. The longest communications had 6 cycles. If the phisher insisted aggressively on personal information, the communications was terminated. From the content of the communications, the attacker's procedures and his argumentation to obtain the required information were primarily examined. A detailed analysis of the texts from the communications aimed to answer the question of whether the phisher is a robot or a person. Further considerations are being made within the team on how to continue researching phishing attacks.

Key words: Phishing email, communications with phisher, fake identity, statistics, analysis.

1. Introduction

Over the past few years, there has been a large increase in targeted cyber-attacks on individuals to obtain sensitive, confidential, or personal information of random or specific individuals or organizations. Depending on how this information is obtained, these attacks can be divided into phishing, spear phishing, and whaling in combination with social engineering. A typical phishing attack is targeted on a large number of people, and the attackers hope to catch a few victims. Typically, a group of people receives an email that looks trustworthy and asks the victim to log on to a website and take a certain action. The email always contains a link to something that looks like a trusted website, and the victim are requested to enter their account details. Once these details are entered, the attacker uses them to steal, defraud, or obtain even more valuable information. The tactic of the attack is to target a larger number of potential victims in the hope of getting some to click on the email and fall for it.

While the goal of phishing is to target a large number of victims, spear phishing targets specific individuals with the goal of obtaining specific information. Prior to a spear phishing attack, the attacker collects publicly available information about the potential victim. They then use the information gathered to persuade the victim to provide more information or to perform a task. As spear phishing targets specific individuals, whaling is even narrower and targets only people in high-level jobs. The social engineering consists mainly of psychological manipulation of people with the aim of revealing sensitive information or performing some action. Social engineering is usually motivated by data theft, malware spread, or political aspects.

Three methods can be used to send phishing and answer processing. First, manual processing, which is strenuous, can contain errors and place high demands on the attacker's attention. The attacker must follow the communication flow and follow up on previous messages. If he or she is inattentive and erring, he or she risks being revealed. Second, attackers can use automatic phishing and automatic responses to the victim's initial response. These messages are prearranged and contain predefined information. Third, the use of bots is now common when communicating with customers of online sales portals or customer support lines. There is no reason for attackers to not use this technology to facilitate their actions. Longer communication is needed to assess whether a person or a bot is behind the attack emails, but the outcome is not always clear cut.

The aim of the article is to describe the results of the research on email phishing attacks against the article's author as a part of the long-term research. The first period in the years 2020-2022 was oriented to the analysis

of phishing emails, and the result was their detailed classification into five segments: Business, Charity, Fund, Transfer, and Others. The Business segment is characterized by the cooperation offered on an investment, contract, or project in the recipient's country, as a realization of a business opportunity. The letter in the Charity segment is mostly written in the Christian spirit, offering the recipient the donation to set up a charity fund and getting a reward for this mission. The alleged sender is an old woman, a widow, her deceased husband left millions of USD. The Fund segment included emails that promised the recipient money obtained from a fund, gift or inheritance. The Transfer segment through email requested cooperation for money or other asset transfer. The Others segment contains emails with marginal significance, not included in other segments, with an unspecific offer, repeated contact, undelivered package, etc. These emails were mostly chosen for the currently described experiment.

The following experiment in the phishing research analyzed defense against phishing attacks by applying the filtration rules of the email client. The rules were set up with the keywords characterizing the individual phishing segment carrying out the selected operation. The phishing emails were transferred into the designated directory, and the defense effectiveness reaches 90%.

The last experiment analyzes the author's communications with phishing attackers. A fake identity was created for the experiment and protection rules were set up. The emails were mostly from the Others segment with nonspecific content; a total of 100 phishing emails were answered. The result of the study is the content of the paper. The literature review confirmed the assumption that no found publication is engaged in a similar attempt, so the results of the research are sure interesting for cybersecurity research.

This paper is organized as follows. The "Introduction" provides an overview of our current phishing research and briefly informs about the latest experiment; following the "Literature review". The research methodology, fake identity, and basic security rules of the experiment are described in the section "Research methodology and security rules". The section "Description of the experiment and its statistical analysis" presents the intention of the experiment and its statistical overview; the section "Detailed analysis of the selected cases" deals with interesting communications with phishing attackers and analyzes them in detail. The "Discussion and conclusions" section summarizes the results of the experiment and suggests possibilities for further research.

2. The literature review

It is increasingly possible to see a combination of social engineering and phishing, where in many cases cyber-attacks are not targeted at the organization as a whole, but at the employees of that organization. In this case, specific human characteristics and psychology are used to circumvent the organization's technical security measures. Social engineering is increasingly becoming a more widespread approach used to compromise individuals and organizations, where it is relatively more advantageous to compromise a person compared to exposing a vulnerability in a security system. Social engineering attacks include physical, social, and technical aspects that are used at different stages of the attack. Even if such an attack is initially unsuccessful, any penetration into individual and organizational security processes can be exploited for future attacks.

The issue of social engineering and phishing is addressed in many publications, often focusing on the general characteristics of social engineering and its detailed characterization, such as A Study on the Psychology of Social Engineering-Based Cyberattacks and Existing Countermeasures (Siddiqi, et al., 2022). The study describes cyber-attacks that are based on social engineering, a process that exploits human psychology. It also provides an in-depth analysis of the approaches used to conduct social engineering-based cyberattacks. This study examines the human vulnerabilities that criminals exploit in security breaches. Further, the study highlights existing approaches, including machine learning-based methods that are being used to combat cyber-attacks.

The human factor as a cybersecurity vulnerability is elaborated on in the paper Vulnerabilities of Cyber Security of Technical Intelligentsia in Relation to Social Engineering (Kononovich, et al., 2021). The paper examines the impact of social engineering on human and collective consciousness, and on the human subconscious. The focus is on the influence of social networks and their functioning on information perception and thinking. The analysis mainly focuses on the vulnerability of educated people and describes recommendations to counter social engineering attacks, in particular: general system recommendations for a system of protection against social engineering attacks; methods of self-defense against the attack of cognitive bias "myside bias"; recommendations of non-system methods and means of protection against social engineering.

A different view of social engineering is presented in an interdisciplinary view of social engineering: a call to action for research (Washo, 2021), which examines the topic of social engineering from an interdisciplinary

perspective. The article reviews literature from the fields of information technology, psychology, and business, explaining the interconnectedness of social engineering with these fields and the need to understand it from multiple perspectives. The literature review is followed by an ethical perspective that analyses social engineering research from a philosophical and professional perspective. The proposed framework provides readers with a flexible model that they can use in their studies with an emphasis on a philosophical or practical ethical perspective.

In the field of social engineering, prevention is equally important in terms of preparing the employees of an organization for possible social engineering attacks. One of the articles that deals with this has been published under the title The Role of Employees' Information Security Awareness on the Intention to Resist Social Engineering (Grassegger and Nedbal, 2021) examines individual and organizational factors that influence employees' awareness of information security and the possibilities to protect against social engineering attacks. The paper evaluates a survey of 136 employees and clearly shows the importance of information security education and awareness. It also highlights the importance of the impact of individual risk-taking behavior on employee awareness of information security.

Another paper (Yeng, et al., 2022) conducts a quantitative and qualitative survey of doctors in the field on simulated phishing attacks using SMS. The results show that doctors who prioritized patient care were not as susceptible to those simulated phishing attacks because they did not pay attention to those messages. The paper also highlights the impact of employee security training on overall security in healthcare facilities. The importance of raising awareness of social engineering threats is discussed in (Alghenaim, et al., 2022), (Singh and Singh, 2022).

During the COVID-19 pandemic, there was a large increase in social engineering attacks. The reason for this increase lies in the increase in the amount of time that users, employees, students, and businesses began to use online environments in a big way. This increase in people's online presence is almost never preceded by education about cybersecurity and the different types of attacks to which the everyday Internet user may be exposed. This situation creates an opportunity for cybercriminals to attack, with social engineering attacks being the most common type of attack in a pandemic. The issue of the increase in cyber-attacks using social engineering during the COVID-19 pandemic has been addressed in the following papers (Venkatesha, et al., 2021), (Li and Lalani, 2020), (Kumaran and Lugani, 2020).

3. Research methodology and security rules

The research methodology in the published study is based on experience from previous research on phishing attacks. The data source for the experiment makes up of phishing emails, captured as a result of a previous experiment using email filtering rules in the MS Outlook client.

The basic security measure of the experiment is the creation of a fake identity - an email account of a non-existent person at gmail.com. Selected emails from the phishing email account were gradually copied to the fake identity email account and sent as a response to the phishing attack.

For communication, typical texts were created in advance, which became the content of the answers. If the phisher responded, communication continued from the fake identity email account. All communication was accurately recorded for the subsequent evaluation of the experiment.

Overview of additional safety measures:

- Do not click on URL addresses from phishing emails.
- Communicate outside the university computer network.
- Do not open suspicious email attachments.
- Do not send personal data (even about fake identity).

4. Description of the experiment and its statistical analysis

The experiment took place between May and July 2022 with the aim of finding out how communication with phishing attackers takes place. A total of 100 phishing emails were answered and their statistical evaluation is in Tab. 1. Significant parameters of the communications are the number of responses and the time it takes for the phisher to respond, see Tab. 2.

In the course of communication, some phishers attached arguments so that their identity and demands could

be considered real and true. Photos of the alleged corresponding person (see Fig. 1), their identification cards (see Fig. 2), and certificates confirming the winning of the fund (see Fig. 3) were sent. The patience of the phishers was sometimes admirable, they were not deterred by any negative response or rejection.

From the experiment, a clear experience can be drawn that it is possible to communicate with phishers without risk, provided that proper security measures are observed. A lot of interesting information was obtained, but the decision on whether a phisher is a person or a robot could not be unequivocally confirmed, see the next section.

Table 1: Email communication statistics

Ord	Email communication	Num	%
1	Without an answer	68	68
2	- of which undelivered	8	8
3	Aswer 1 times	10	10
4	Aswer 2 times	8	8
5	Aswer 3 times	11	11
6	Aswer 4 times	1	1
7	Aswer 5 times	1	1
8	Aswer 6 times	1	1
9	Total communicating	32	32
10	Total answers	74	

Table 2: Speed of responses

Ord	Speed of response	Num	%
1	On the same day	13	18
2	In 1 day	12	16
3	In 2 days	6	8
4	In 3 days	15	20
5	In several days	28	38

Figure 1: Example of photo of the alleged corresponding person

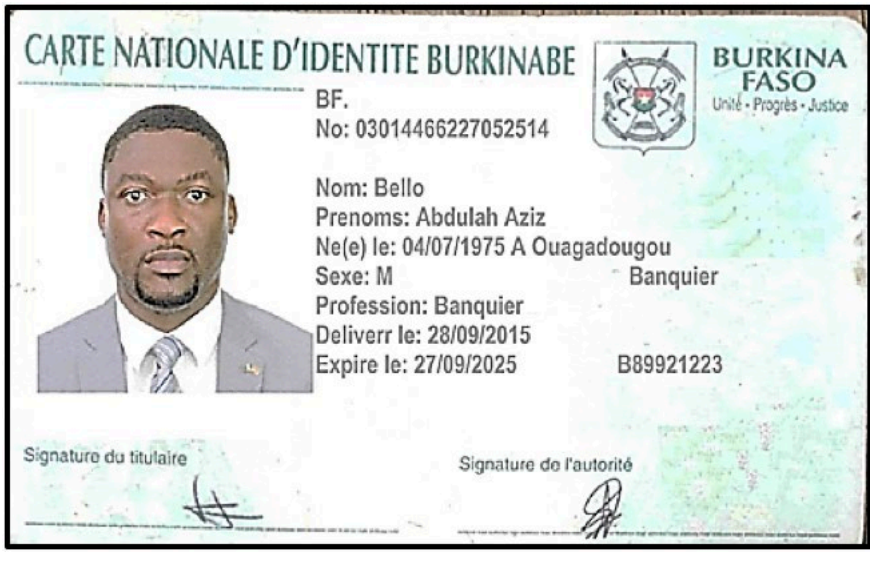


Figure 2: Example of ID-card of the alleged corresponding person

Figure 3: Example of a certificate

5. Detailed analysis of the selected cases

In the section, the analysis of selected communications is provided. The three longest communication is taken into consideration, analyzed, and conclusions drawn. The longest communication has been six rounds before being terminated due to the victim's decision. The main topic is the fund offer from a terminally ill woman. The second communication lasts five rounds, and the topic is a fund offer to help the poor. It is finished by the victim's decision, also. The third communication has four rounds. An attacker acts as an attorney to propose a business offer.

5.1 An offer from a terminally ill person

The longest communication lasted six rounds. The initial phishing email is short, the attacker poses as a terminally ill woman and asks the victim for help in transferring $5.5 million. 30% of this amount can be kept by the victim and 70% is to be used to help the poor.

After the first response from the victim, a comprehensive email is received. There is an emphasis on the urgency of the situation (3x word immediate, 2x word urgent) due to the condition of the patient. Religion is a very distinctive feature of email. The attacker writes about the Bible and about the service to God; the word God appears here 4 times. The reason for contacting the victim is God's will. This is followed by an assurance of the legality of the transfer and a request to send personal information (full name, nationality, occupation, age, picture, mobile phone), which will be used to write the authorization letter for the benefit of the victim. Compassion for orphans and the less privileged is highlighted several times.

After a negative response from the victim who does not want to share any personal data, an even more comprehensive email is received. The attacker again stresses the urgency of the situation, charity, and God's will. The attacker writes that she trusts the victim and would like to get to know the victim better, so she asks for personal information. There is a contradiction here: can one trust someone you do not know? Furthermore, detailed banking information is attached, and a fund release form is attached. The mentioned African Development Bank and its branch in Burkina Faso exist and the international phone code is also correct.

However, the address and phone numbers do not match the official contact details of the branch. The attacker's contact email adb-bf11@accountant.com does not appear to be fraudulent at first glance. Nevertheless, the domain accountant.com is not the real domain of the ADB bank, the real domain is afdb.org.

The victim replies that he cannot make the transaction. The attacker's response is very short, the victim is supposed to follow the bank's instructions. After nearly 2 weeks, the attacker asks if the victim has contacted the bank. The victim responds negatively and expresses concern about the legality of the transaction. The attacker reassures the victim and asks again if the victim has contacted the bank. The victim again expresses concern. The attacker responds after 12 days, reassures the victim again, and insists on contacting the bank. The victim stops the communication here.

The attacker communicates between 5:53 and 18:28 and the time between replies varies between 8 hours and 12 days. Long response times conflict with the urgency that the attacker is trying to evoke in the victim. After the initial response of the victim, automated responses may have been used. There is a general salutation in the response, and nothing from the victim's previous email is used. On the other hand, the response was sent 7 days late, suggesting manual processing. The attacker's motivation to obtain information is at a high level. The attacker responds logically, has an overview, and refers to previous communications. Salutations and goodbyes change slightly, which also indicates manual processing. Longer delays can be caused by the large number of emails being processed.

Language is at a very good level in the initial phase of communication. Later, the language level decreases and contains minor errors. There is a clear transition from the prepared parts of the attack to the unprepared parts, where the attacker's improvisation is needed when communicating with the victim.

5.2 Fund offer for the poor

Delivered phishing contains an offer of funds to help the poor and raise living standards in the recipient's region. The recipient was selected based on a profile from unspecified Microsoft and Google lists. The attacker requests a replay with name, address, country, age, sex, occupation, and phone number. Some of the data should not be requested by the attackers, as it should be known to them from the mentioned profile from Microsoft or Google lists. This makes the email look suspicious at first glance.

Although the victim answers negatively and does not want to share any personal data, the attacker continues his attempt and informs the victim about the funds deposited in the bank ($650,000.00), again requesting a name, address, and phone number. The attacker mentions a 'reference number' and encourages contacting the bank to gain access to the money, which is most likely to increase the credibility of phishing. However, he mentions informationschasebank@gmail.com as the bank's contact email. The address is very suspicious because the gmail.com domain is not used by banks, but they register their own domains. After the victim questions how to access the funds, the attacker responds illogically. In reply, he writes only that the money in the bank is waiting for the transfer. It is only after repeated questioning of the victim that the attacker replies that a fee of 100 euros is needed to transfer the money, and then the money will be transferred immediately. Interestingly, the attacker does not specify or require an account number where the fee or funds should be transferred. The victim responds negatively, no money will be sent, and expresses concern that it is a fraud. The attacker replies again illogically, using the same formulated answer that the money in the bank is waiting to be transferred. After re-emailing information that the victim will not send any money and fears an attack, the attacker sends the same reply for the third time about the waiting money in the bank. The victim terminates the communication.

It is clear from the communication that this is not a professional group, but amateurs. Errors can be considered to require information that the attacker should know, use of a non-bank domain, illogic in replies, and repeated replies. These are errors of a fundamental nature and it is obvious at first glance that this is a fraud. Partial automation is not used here, where after the initial reaction the victim can be automatically sent an email with information about the deposit in the bank, the reference number, and the contact to the representative of the bank. Several-day time delays do not correspond to automation, but rather to manual processing. The time interval of the attacker's responses from 9:37 to 15:40 also corresponds to manual processing. Phisher’s motivation to obtain personal data is not high, due to non-consecutive communication and longtime delays. However, it is possible that the long response time is caused by the large number of emails that the attacker has to manually process. It is also possible that this communication was provided by a robot that had learned the answers and did not modify them in any way. In this case, the time delays could be set deliberately or could be due to a large number of emails. The language level of the attack emails is good. However, it cannot be concluded

that the attackers are linguistically educated, as even the free English translators are at a very good level currently. The constant level of English could also be explained by the use of bots in communication.

5.3 Mutual business proposal

Initialization phishing is very short, the attacker offers mutual business and gives an email address for replies. It is suspicious that the given contact address is different from the sender's one. After a positive response from the victim, a message is sent with detailed information about the business offer. The attacker acts as an attorney for the deceased Mr. Malik, who is looking for his offspring for the settlement of the estate of $23,600,000.00. It is not important that the victim is not biologically related, it is enough that he has the same surname. Unfortunately, this is the last attempt to settle the estate before the entire fund is confiscated. The attacker states that he will split the amount into a 50% ratio for him and 50% for the victim. The attacker wants the victim to act as a relative and the attacker will use his position as a personal attorney for the smooth progress of the transaction. At the end of the email, it is stated that everything is legal. At this point, it is obvious that this is a financial fraud offer. The attacker uses the same surname as the deceased person and the victim, which is supposed to increase credibility and the success of the attack. So far, the attacker does not require any personal data or any information; he just ascertains the possible interest. He also does not induce any time pressure.
A positive response from the victim is followed by a comprehensive email repeating the information from the previous communication, and, in case of interest, the attacker offers the option of filling out a form to release funds and handle other needed documents. Therefore, he requires sending personal information to confirm the identity of the victim (international passport or driver’s license or any valid ID), contact address, occupation, marital status, mobile number, landline, and fax number). To increase credibility, the attacker attaches to the message a notification from the bank about the last chance to claim the funds of late client Engr. Phillip Malik. The attacker creates a slight time pressure.

After a negative response from the victim and a reluctance to share personal information, the attacker responds with pleas, appeals to the partnership, and assurances about the legality of the transaction. He gives the victim his word that he will not misuse personal information and again asks for the required data to be sent. To promote confidence, the attacker encloses a copy of the passport. According to the date of birth on the passport copy, the person is 60 years old, while the passport photo shows a young man. Here, too, the attacker makes another mistake and signs the email with a different name. Again, the victim responds negatively, and the attacker responds with more pleas and assurances. The victim terminates the communication.

The time span of the attacker's replies ranges from 12:27 to 20:15. It could be working hours in a different time zone. However, such a zone does not correspond to the attacker's declared Spain as his place of life. The delay between replies varies from 2 to 9 days. Initially, the attacker responds within 2 days; later, with the victim's unwillingness to share data, the attacker's motivation decreases, and his replies are sent with greater delays. The attacker responds logically, orients himself in the communication, and refers to the previous communication. In the later stages of communication, the attacker makes mistakes. This suggests manual response processing.

The language is at a very good level throughout the communication. Invariable addresses, constant English style, and similar repetitive passages could also be explained by the use of bots in communication. Also, the initialization of phishing and the response to the victim's first positive action could have been processed by automation or bots. Taking the surname and embedding it in the text of the email does not require any activity of a human operator. The parts of the communication where errors occurred are likely to have been processed manually.

6. Discussion and conclusions

Automatic phishing and bots can provide the initial stages of communication. Errors occur in the later stages of communication, where attackers need to react and improvise. This can indicate the limited scope of bots and the manual processing of emails. Bots can operate nonstop, whereas human operators only react at certain time intervals. Thus, it can be deduced from the times of communication whether it could have been a bot or a human. Language level and text style can also be a sign of automatic processing.

In the early stages of communication, where texts and information are defined, automatic responses or bots can be used, using online English language translators, which are currently at a very high level. In subsequent parts of communication, where there is a clear change in style and where errors occur, it is likely that the response

has been processed by a human. From the above, it can be deduced that such indicators can be used to determine whether a human or a machine provides communication.

Further research should be oriented to the best evaluation and use of all the results and conclusions to date for developing defensive abilities and skills of users against phishing attacks. The goal will be both high-quality teaching and training programs, as well as defense methodologies and rules.

Acknowledgments

This article presents the results of the research in cyber security of the Department of Informatics and Cyber Operations, University of Defence, as part of the project KYBERSILY (DZRO-209, 2022).

References

Alghenaim, M. F., Azaliah, N., and Rahim, F. A. (2022). "Exploring the Factors Influencing Employee Awareness of Social Engineering Threats: A Review". Applied Mathematics & Information Sciences, July, pp. 491-500.

DZRO-209. (2022). KYBERSILY: "Cyber forces and assets," Brno, University of Defence, 2021-2025.

Grassegger, T., Nedbal, D. (2021). "The Role of Employees' Information Security Awareness on the Intention to Resist Social Engineering". Procedia Computer Science, 181, pp. 59-66.

Kononovich, V., Kononovych, I., and Shvets, O. (2021). "Vulnerabilities of Cyber Security of Technical Intelligentsia in Vulnerabilities of Cyber Security of Technical Intelligentsia in. Odesa", ISIT 2021, pp. 127-136.

Kumaran, N., Lugani, S. (2020). "Protecting businesses against cyber threats during COVID-19 and beyond". [Online] Available at: https://cloud.google.com/blog/products/identity-security/protecting-against-cyber-threats-during-covid-19-and-beyond.

Li, C., Lalani, F. (2020). "The COVID-19 pandemic has changed education forever. This is how". [Online] Available at: https://www.weforum.org/agenda/2020/04/coronavirus-education-global-covid19-online-digital-learning/

Siddiqi, M. A., Pak, W., and Siddiqi, M. A. (2022). "A Study on the Psychology of Social Engineering-Based Cyberattacks and Existing Countermeasures". MDPI Journals.

Singh, I., Singh, Y. (2022). "Cyber-Security Knowledge and Practice of Nurses in Private Hospitals in Northern Durban, Kwazulu-Natal". Journal of Theoretical and Applied Information Technology, 15 January, pp. 246-267.

Venkatesha, S., Reddy, R. K., and Chandavarkar, B. (2021). SN Comput Sci., 6 February.

Washo, A. H. (2021). "An interdisciplinary view of social engineering: A call to action for research", Amsterdam: Elsevier.

Yeng, P. K., Fauzi, M. A., Yang, B., and Nimbe, P. (2022)." Investigation into Phishing Risk Behaviour among Healthcare Staff!. MDPI journals.

Agile Methods For Improved Cyber Operations Planning

Dr. Jami M. Carroll
Unaffiliated, Somerset, MA, USA

[jcarroll@prisidian.com]

Abstract: Cyber Ranges provide an interactive simulated environment of hardware and software for simulation. This closed environment provides a safe and legal environment where cyber warfighters can refine their skills. They enable mock cyber mission rehearsal of operation playbooks. Simulated cyber capabilities in the cyber range parallel the intelligence, surveillance, and reconnaissance (ISR), Order of Battle (OOB), and battle damage assessment (BDA) in a closed, safe environment for experimentation. Scrum has been used in collegial cyber competitions with success because it has allowed Capture-the-Flag cyber games to create quicker simulations. Defense Innovation Units (DIUs) are using agile Scrum processes to numerous warfighting areas in order to make them more agile. This research argues that the agile software development processes could be used to optimize the planning and execution of offensive, defensive, and operation and maintenance (O&M) of cyber warfare simulations within cyber ranges. O&M can be done quicker, new exploitable modules can be includer more rapidly, and the capability can be reconstituted to the appropriate skill level for the next set of trainees quicker. The White team as maintainers of the networks, systems, applications and cyber tools select the CVE exploits and spend an enormous amount of time installing and configuring these capabilities for the next set of trainees. Quite often, there are different skill levels which require multiple builds and the ability to refresh the cyber range with varying levels of cyber trainee complexity. This requirement to restore the cyber range quickly with a variety of builds, varying levels of difficulty, and ensure the experiential learning is maximized with the best availability lends to agile methods such as Scrum could lend to improvements with cyber operations. This research will illustrate how a cyber range could leverage agile Scrum processes to provide an improved cyber range environment quicker and with more capabilities.

Keywords: Cyber Range, Cyber Security, War Game, Agile Methods, Scrum

1. Introduction

Disclaimer: All statements of fact, opinion or analysis expressed are those of the author. The views and opinions expressed herein by the author do not represent the official policies or positions of the United States (U.S.) Department of Defense (DoD), U.S. Navy, or other agencies or departments of the U.S. government and are solely representative of the views of the author. This does not constitute an official release of DoD or Navy information.

This study will examine how Agile Methods like Scrum can be applied to Cyber Ranges to significantly improve the offensive and defensive cybersecurity skills of the red and blue teams while providing a more responsive cyber range for the white team to manage (Vykopal, Jan et al., 2017). National Institute of Standards and Technology (NIST) National Initiative for Cybersecurity Education (NICE) describes cyber ranges as cloud-based environments allowing educators, government, industry, and the military to provide offensive and defensive cyber security training to participants to expand their skills (NIST NICE, 2018). The U.S. Cyber Range indicates that these simulated representations of networks, systems, applications and cyber tools that provides performance-based learning and assessment; develop cyber teamwork; enable real-time feedback to be given in an environment that simulates real-world scenarios while operating in a safe environment where new tools, techniques, and procedures can be applied (U.S. Cyber Range, n.d.). Red team are the offensive cyber attackers trying to breach the networks, systems, and applications using techniques ranging from hacktivism to nation-state level attacks. Blue team are the defensive cyber defenders trying to protect the networks, systems, and applications. The White team designs the cyber war game, injects stimuli such as scripted attacks, simulated intelligence reports, and tasking orders requiring red and white team responses (Sullivan et al., 2018). Lectures and labs alone provide a limited facet of how the cyber warfighter will need to respond. In actual war, there are military Operation Plan (OPLAN) which lays out how they will fight the battle, including the cyber battle. A key element is Intelligence Preparation of the Operational Environment (IPOE) which will focus on developing course of actions (COAs) to support the warfighter. This IPOE will include cyber elements related to Intelligence, Surveillance, and Reconnaissance (ISR), Order of Battle (OOB), and the Battle Damage Assessment (BDA) as they are executed (Couretas, Jerry, 2022). Cyber Range warfighters will use these battle scenarios to develop Tactics, Techniques and Procedures (TTP) that can lead to improved tradecraft in handling real world responses when not in the cyber range. Tradecraft in cyber warfare requires extraordinary secrecy, deception, stealth and strategy to hide the exploit while using intelligence, planning and operations to achieve the attack before the defender can patch, reconfigure, or use other countermeasures to thwart the attack requiring the attacker to change their cyber operations attack plans (Gartzke & Lindsay, 2017). Because the operation of the cyber range is so important to the learners within the constrained training period, it is critical that the Cyber Range is constantly available for use. Unlike a typical enterprise information technology (IT) system, the cyber range is

constantly going through an exercise lifecycle covering the exercise's development, dry run, execution, evaluation, and repetition as well as coming up with new scenarios as vignettes within the OPLAN (Vykopal, Jan et al., 2017). In 2022, MITRE published 25,068 new vulnerabilities in the Common Vulnerability Exposure (CVE) database – an increase of 24.3% beyond 2021 (Jacobs et al., 2023). The White team as maintainers of the networks, systems, applications and cyber tools select the CVE exploits and spend an enormous amount of time installing and configuring these capabilities for the next set of trainees. Quite often, there are different skill levels which require multiple builds and the ability to refresh the cyber range with varying levels of cyber trainee complexity. This requirement to restore the cyber range quickly with a variety of builds, varying levels of difficulty, and ensure the experiential learning is maximized with the best availability lends to agile methods such as Scrum could lend to improvements with cyber operations. The primary research question is: How can agile software development processes like scrum be applied to improve Cyber Range operations and learning? It is anticipated that the results of this study could be used to enhance all Cyber Ranges and that this foundational work may help U.S. Service/Agencies, Allies, and Partners.

2. The Evolution of Cyber Ranges

The first cyber ranges were largely classified Department of Defense (DoD) built to test cyber as weapons much like weapon testing grounds used for the first nuclear weapons. These early DoD cyber ranges were used by cleared government and contractors much like going to the gun range. Because a DoD clearance, classified networks and need-to-know / need-to-use were involved, very few users had access to these cyber ranges. The Michigan Cyber Range is considered one of the first non-government cyber ranges created in 2012 after discussion with leadership at the White House and U.S. Department of Homeland Security (DHS) by the Michigan state cybersecurity team. Michigan focused on two goals with their Michigan Cyber Range: 1) workforce development and skills training and 2) development by academic and university organizations to replicate the U.S. Military Academy at West Point Information Warfare lab (IWAR) (Lohrmann, Daniel, 2018).

Another expansion towards the needs for cyber ranges evolved as a result of the National Security Agency (NSA) creating the Centers of Academic Excellence (CAE) in Information Assurance Education degree program created in 1999. The University of Texas at San Antonio (UTSA) is credited with starting the first U.S. collegial cyber competitions in the southwest in 2005 to extend these NSA degree programs with applied learning. In 2010, the U.S. House passed Resolution 1244 recognizing the importance of cybersecurity education programs and the need for cyber ranges and cyber exercises and the need for a National Collegiate Cyber Defense Competition (NCCDC) to promoted cybersecurity curriculums (Schafer, Joseph H., Morrell, Chris and Blaine, Ray, 2022).

As these NCCDCs started identifying their learning outcomes, they realized there were three main groups of learning. One group, the defenders, were responsible for designing, configuring, and protecting the network and its resources. A second group, the attackers, were responsible for conducting the vulnerability assessment / penetration testing of the network infrastructure, databases, mail servers and Web servers. A third group was responsible for the day-to-day operation and maintenance of this environment and the ability to quickly restore it after exercise completion or worse yet, the addition of new attack/defend capabilities, or even catastrophic failure resulting from attacks. NCCDCs came to realize that maintaining an excellent learning environment with a high degree of availability for the learning teams was a daunting task.

Cyber ranges train cyber warfighters against cyberthreats by criminals, state actors, and non-actors. In order to provide the necessary offensive and defensive cybersecurity skills, cyber ranges provide a critical capability for active learning where theory is supplemented with application (Vykopal, Jan et al., 2017). Lectures and labs alone only provide one side of the learning and lack the true impact gained from analyzing how each side affected the attack. The combination of combining offensive and defensive training allows interactions between the offensive and defensive players inside a safe and legal closed environment that closely mirrors real world operations. (Ferguson, Bernard, Anne Tall, and Denise Olsen, 2014; National Initiative for Cybersecurity Education (NICE), 2018). These virtualized environments have Web servers, databases, multi-tiered architecture, and network infrastructures that are similar to real-world operations (Vykopal, Jan et al., 2017).

Vulnerability assessment is the initial finding of potential vulnerabilities that may be exploitable while penetration testing is the actual exploitation of the vulnerability found; the actual exploitation of the system crosses the line of potentially illegal acts similar to espionage that occur in a cyber range. The sequence of steps such as footprinting, fingerprinting, enumeration, research, escalation, repeat visits, and covering of tracks cover the steps involved in vulnerability assessment through penetration testing to achieve the attack while maintaining stealth. Figure 1 illustrates the typical steps used in vulnerability and penetration testing of systems.

Vulnerability assessment involved the collection of information through footprinting, fingerprinting, enumerating, and then researching the possible attack vectors. Once the vulnerability phase is completed, the penetration tester develops a list of possible attack vectors based on either ease of attack, effects achieved by the attack, or when simulating a hacker or state actor, acquiring a particular data element. Acquiring a particular data element in the NCCDC is sometimes referred to a "capturing the flag" (Carroll, 2017).

	Common Approaches	Description	Comments
Vulnerability Assessment	Footprinting	Where do the IP addresses/systems show up?	• Reconnaissance (active (scanning) or passive (sniffing) • External – DNS and external-facing systems (nslookup, whois, dig, SamSpade, NMAP, & Zone Transfer) • Internal – SNMP, Unix, Windows, LDAP, SMTP, & NMAP) *Note: Approach is slightly different if coming from inside the network than from outside the network*
Vulnerability Assessment	Fingerprinting	What are the ports and services for the OSs and Applications?	Fingerprinting (scanner with OS recognition)
Vulnerability Assessment	Enumeration	What version of a service exists? (Example – banner grabbing of a web server or mail server)	Enumeration, potential vulnerabilities
Vulnerability Assessment	Research	Common Vulnerability Database (CVE.mitre.org); National Vulnerability Database (http://nvd.nist.gov/);Open Source Vulnerability Database (www.osvdb.org)	
Penetration Test	Escalation	Privilege escalation, Denial of Service, Man in the Middle	Obtain access
Penetration Test	Repeat visits	Backdoors - Hackers do this - Pen Testers simulate unless authorized to perform destructive testing (DoD) - Computer Network Attack (CNA)/Computer Network Exploitation (CNE) may do this when authorized	Maintain Access
Penetration Test	Covering tracks	Log zappers, log stoppers - Hackers do this - Pen Testers simulate unless authorized to perform destructive testing (DoD) - Computer Network Attack (CNA)/Computer Network Exploitation (CNE) may do this when authorized	Erase evidence

Figure 1. Typical Steps Used in Vulnerability and Penetration Testing of Systems

3. Cyber Wargaming: A Venue for Mission Rehearsal

War games attempt to simulate military operations against an opposing force in what might be similar conditions. The concept of war games goes back to Sun Tzu's 5th century BC era and the term was first coined in 1824. The advantage of the wargaming experience for military leaders is acquiring a lot of practical experience without "real-world penalties" (Sullivan et al., 2018, p. 92). Cyber wargaming simulates the "red team" as the attackers, the "blue team" as the defenders, and the "white team" as the designers and executors of the war game (Sullivan, D.T., et al., 2018). The white team designs the war game with a pre-planned state and objectives for the red and blue team to achieve. Many of these cyber ranges have added gamification objectives where quantitative points, levels, and leaderboards are gained by the team based on their offensive or defensive efforts (Diakoumakos, 2021). Gamification has increased student interest / enthusiasm because the game scenarios very closely emulate real-world situations within virtualized environments in cyber ranges with practical training gained through experiential learning at educational institutions. With a lack of cyber warfighters and an ever increasing number and sophistication of attacks, a continuous stream of cybersecurity education can be offered (Jelo, M. and Helebrandt, P., 2022).

Cyber Ranges enable mock cyber mission rehearsal of operation playbooks (Vykopal, Jan, 2017). Mission Rehearsal Exercises (MREs) are conducted prior to many major operations. Prior to the infamous Operation Neptune Spear which led to the raid on Osama bin Ladin's complex and his death, numerous mockup MREs were conducted to simulate the operation prior to execution (Gratch, Jonathan and Marsella, Stacy, 2003; Christensen, Gitte Højstrup, April 2017; Kiras, James D., 2017).

For the military, these isolated environments enable cyber missions to be rehearsed, and capabilities evaluated. Much like in actual war, there is the equivalent of a military Operation Plan (OPLAN) which lays out how they will fight the cyber battle based on hypothetical scenarios, military objectives, and military doctrine for the forces being used. Cyber OPLANs will have wargaming segments supporting cyber intelligence surveillance reconnaissance (ISR), creating a cyber-oriented Order of Battle (OOB), and a cyber intelligence battle damage assessment (BDA) that are broke into smaller sections of ISR, OOB, and BDA that are handled by cyber teams

from a variety of military and government entities with different skill sets (Blasch, Erik, and Micheline Bélanger, 2016).

In order to keep pace with technology, since the early 2002s, quite a bit of the U.S.'s Command, Control, Communications, Computers and Intelligence (C4I) systems are built with standard Commercial-Off-the-Shelf (COTS) hardware, software, and firmware and far less Government-Off-the-Shelf (GOTS) hardware, software, and firmware. This allows technology insertion to happen quicker and at a significantly decreased cost. Another advantage is that many of the standard information technology capabilities found in industry are also directly used in government C4I systems (Huskey, T.W., 2007). This allows new technology in industry to be used in cyber ranges and provide a similar environment to what cyber offensive and defensive teams can be trained within the cyber range (Couretas, J.M., 2022).

4. Agile Software Development Processes

Since cyber ranges must create multiple simulated representations of networks, systems, applications and cyber tools based on: 1) varying student skill level of students being trained, 2) quickly restore these representations to a pristine environment for the next set of student, and 3) constantly adding new real-world vulnerabilities, DevSecOps could significantly improve how quickly cyber ranges could be setup. The quicker the White team can integrate new vulnerabilities and simulated injects, the closer to a real-world environment that the cyber range will take on. The White team can simulate simulated capabilities as models that allow attackers to create exploits while defenders can protect. Examples of these models include code injection attacks, denial-of-service (DoS) attacks, identity-based attack, insider threats, malware, phishing, spoofing and supply chain attacks. These models provide valuable learning opportunities for the Red team to perform exploits while the Blue team hardens their systems against these exploits. The counterbalance between the attackers and defenders leads to the development of TTP that is highly applicable to real-world operations. Agile methods using Scrum have made up to a fourfold improvement compared to non-Agile methods (Sutherland & Sutherland, 2014).

Traditional software development lifecycle (SDLC) approaches like the spiral or waterfall models can take several years to decades for complex systems. These traditional SDLC approaches often lack the design thinking & agile approaches to meet cutting edge capabilities (Pereira, J.C. and de FSM Russo, R., 2018). Worse yet, long development cycles can lead to endless requirement creep, unfulfilled system end state, cancelled program, and hardware refresh prior to delivery since the software evolves faster than the hardware (Giachetti, R.E. and Van Bossuyt, D.L., 2022). Iterative and incremental evolved between the 1950s and 1970s with many lightweight agile software approaches gaining maturity in the 1990s. The best of these agile approaches came together when the Agile Alliance developed the Agile Manifesto in 2001 (Fowler, M. and Highsmith, J., 2001).

Highsmith and Cockburn argued that software development metrics indicated that most software development using traditional software lifecycle development methodologies failed to satisfy the customers. The main reason it failed was because of a focus on conforming to a plan as the primary goal and satisfying the customers at time of delivery as a secondary goal. In addition, requirements, scope and the technology used was constantly changing throughout the projects period of performance -- this further meant that the final delivery to the customers was even further away from desired. The assumption under traditional software lifecycle development methodologies was that any variation from the originally devised plan would cause unacceptable errors which would further cause deviation from the initially designed plan (Highsmith, J. and Cockburn, A., 2001).

Agile software development processes outlined four tenets with the goal of improving software development and reliability while providing significantly better "customer value". These tenets desired the following outcomes: 1) focus on the individuals and their interactions as being more important than processes and tools; 2) develop working software that fulfilled its intended purpose over extensive documentation; 3) collaborate directly with customers over negotiation via contracts; and 4) respond to necessary changes to the software over following a hard development plan. The Agile Manifesto developed 12 principles focused on early-on and continuous software development that provided value-added software in less time and of a greater quality and reliability (Fowler, M. and Highsmith, J., 2001). Sutherland & Sutherland argued that Agile software development processes can help developers quadruple their software delivery when Agile is fully embraced (Sutherland, J., & Sutherland, J. J. (2014). Agile software development processes accepts that while errors can occur, they cannot be completely eliminated and rather than putting so much effort in eliminating rework, the approach should be to reduce costs associated with this rework, but continue to focus on software quality. Agile software development processes also accepted that it is important to continuously understand the customer's needs for

the software and accept requirements can change, but by staying continuously engaged with the customer throughout short one to four week periods of development in a sprint, the software was more likely to meet the customer's needs. In order for this to work, the most commonly used agile software development process, Scrum, focused on 15-minute daily Scrum team meetings and 30-minute comprehensive Scrum team reviews at the end of each one- to four-week Sprint cycle (Highsmith, J. and Cockburn, A., 2001).

Scrum is an agile project methodology approach that helps a team structure and manage their work in small increments called sprints where the Scrum team applies empirical processes to provide iterative capabilities by ensuring their processes are transparent, inspectable and adaptable as these new increments of change are required (Scrum Org, n.d.). The U.S. military has adopted Agile Processes outside of software development for a variety of warfighting use cases because it achieves results significantly faster than waterfall methodology approaches; the main reason is that Scrum is value-oriented vice Waterfall methodology being schedule-oriented. Uses outside of software development include reorganizing "command and control, quick reaction capabilities, rapid acquisition offices, and US force transformation" (Tomar, A.K., 2017, p. 11). Douglas claimed that Agile Multi-Domain Command and Control is leveraging agile processes to increase advances in how artificial intelligence (AI), autonomous systems, aircraft and nuclear weapons are so critical to military capabilities that Agile processes have been adapted to achieve best possible benefits (Douglas, O.C., 2020). The hardware design for the CubeSat Multi-Mission Bus Demonstration (MBD) project leveraged Agile processes (Huang, P.M. et al., 2012). A major multinational military aviation company uses Agile processes to develop their family of aircraft (Freitas et al., 2020). The Defense Innovation Units (DIUs) like NavalX are focusing on how to provide Agile processes to a variety of warfighting areas, where Scrum can be applied in a militaristic construct (NavalX, 2022). Some of the NCCDCs have started looking at leveraging Scrum in their cyber competition environment because of the operation and maintenance required by the White Team to be able to provide a ready environment for a continuous stream of NCCDC participant teams using their cyber range environment (Novak, H., 2013).

Figure 2 illustrates the Scrum Framework. The following describes the main blocks within the Scrum Framework (Scrum.org. 2023):

- The product backlog lists all functionalities, the requirement for the functionality, a short description of those functionalities. The product owner maintains this product backlog of all desired capabilities for the system. For a cyber range, this may include new capabilities that the offensive and defensive warfighters will get such as the latest technology, technology with known vulnerabilities (e.g. log4J), newer cyber tools being released, or application / operating system patches.
- Sprint planning is done by the product owner, development team, and Scrum Master. It identifies the capabilities that will be done during the one- to four-week sprint. Two hours per number of weeks is allocated to sprint planning. For a cyber range, this may include patching or upgrading software items that were too easy for the offensive warfighters to attack. For the defensive warfighters, it may be applying DoD or business preferred hardening techniques to make the component harder to attack.
- The Sprint backlog is what is being worked on during the one- to 4-week sprint. The Spring backlog must identify the "definition of done" for the capability under development – this becomes their mantra for success. Because these changes are being done in a development environment before pushed to a production environment where the offensive and defensive warfighters are wargaming, both development and production environments can be run simultaneously to allow continuous training events.
- The single Scrum team is a team of less than 10 people focused on development of the Sprint. They will meet daily for no more than 15 minutes to discuss the user story of what the capability is doing, what is and what is not working, and changes that must be done to accomplish the work.
- The Product Increment is the incremental / iterative increment of the product that is either useful or provides value to the product under development.
- The Sprint Review is attended by the product owner, Scrum Master, development team, and all stakeholders / customers / sponsors. It is used to demonstrate the working and tested Product Increment created during the Sprint. This meeting is timeboxed to no longer than one hour times each week of Sprint.
- The Sprint Retrospective is attended by the product owner, Scrum Master, and development team to talk about what went well and what did not go well. It is focused on continuous improvement. This meeting is timeboxed to no longer than one hour times each week of Sprint.

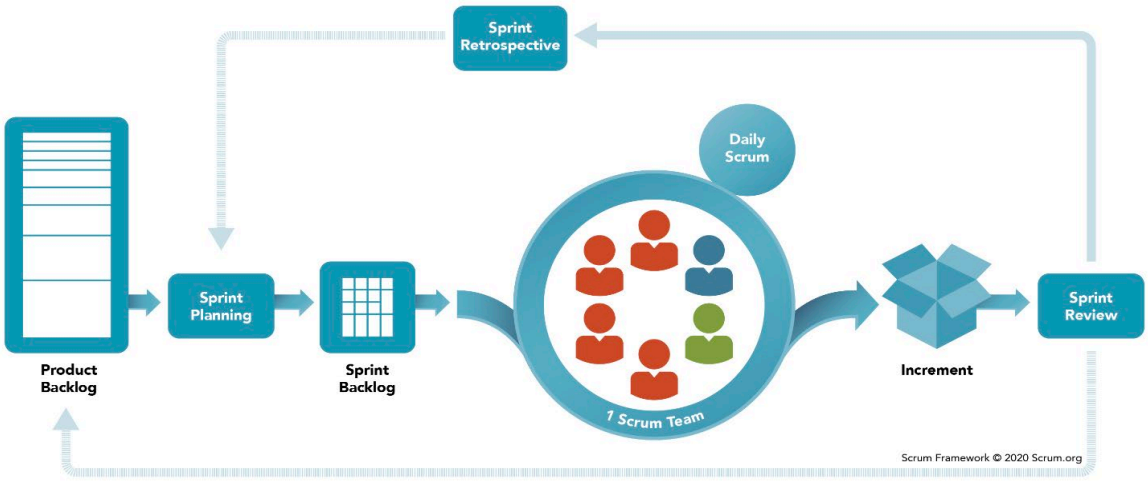


Figure 2. The Scrum Framework.

5. DevSecOps

Development, Security, and Operations (DevSecOps) is a combination of automation, platform and culture that is applied within an information technology (IT) lifecycle. DevSecOps allows rapid and frequent development cycles. Figure 3 illustrates a DevSecOps environment, there is Continuous Integration (CI) coupled with Continuous Deployment (CD) with a Blue Green Pipeline. At the far right, the current production capability is called the Blue Pipeline and represents the current cyber range configuration and referred to as "Current Version N". The next version of the cyber range configuration is called the Green Pipeline and referred to as "Current Version N + 1". The N + 1 version is the one that has received the Scrum Increment. Starting at the far left in Step 1, the Scrum Team is working on the Increment assigned in Sprint Backlog. As the Increment is being worked on, the code branch is put into a repository (e.g. Github) during step 2. During step 3, the code branch is pushed to the Static Application Security Tools (SAST) and Dynamic Application Security Tools (DAST) in step 4 for regression / security testing. If they pass all tests successfully, they are sent to step 5. If they do not pass all step 4 testing, they are sent back to step 1 for rework and continuation with the process. Once the code branch successfully gets to step 5, the software is staged for final acceptance / quality assurance testing. If it passes, it is sent to the N + 1, New Version as a future production release. A package / artifact manager releases the product to the Production environment when it is ready to be fielded.

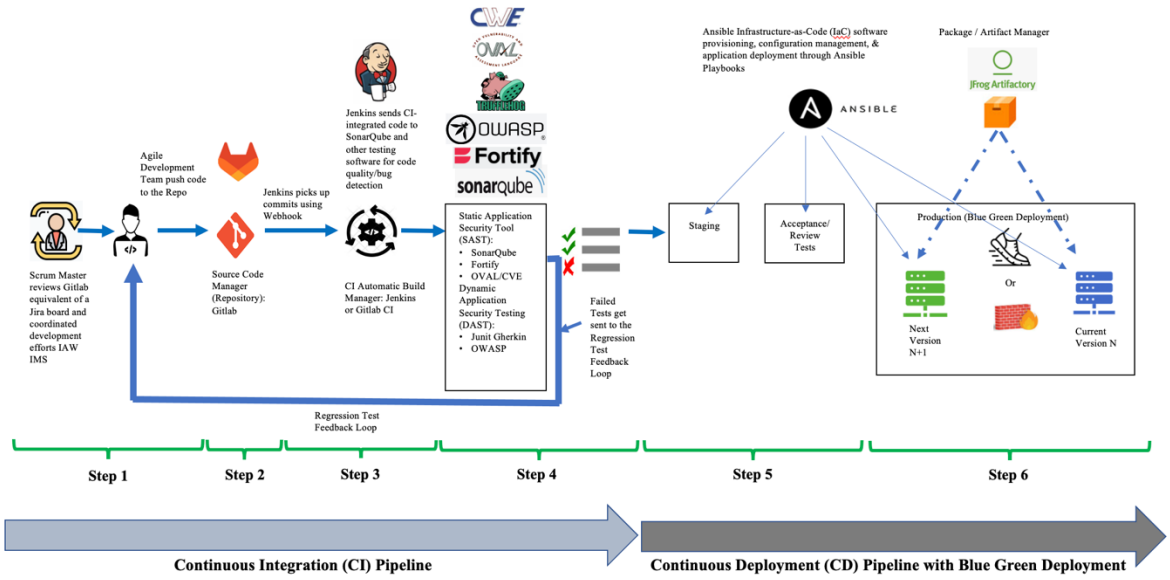


Figure 3. Development, Security, and Operations (DSO) environment

In a cyber range, offensive security testing often requires a complete re-installation or re-imaging of the systems after each training event because the systems have different parameters than when the cyber warfighters started their practice. This is required because software is broken, misconfigured, and services are modified during the penetration testing. Virtual Machine (VM) imaging is one of the best ways to restore the system to the original state before the next red and blue teams can start using them. Since the cyber range is always adding new features, providing timely patches, and security hardening the system, the DSO environment with Blue Green Deployment along with VM imaging allows the White Team an ability to maintain the current and next system for continuous availability of a cyber range suite for classes as they are needed.

6. Conclusion

This research argues that the agile software development processes, such as Scrum coupled with a DevSecOps could significantly optimize the planning and execution of cyber ranges. New capabilities that the offensive and defensive cyber warfighters will get such as the latest technology, technology with known vulnerabilities, newer cyber tools being released, or application / operating system patches could be managed via Scrum and pushed out via the Development, Security, and Operations (DevSecOps) environment extremely fast while providing an environment that allows for the next version to be quickly created and ready to be pushed out as a VM image. Having highly available systems that can quickly support mock cyber mission rehearsal of operation playbooks allows Cyber OPLANs with wargaming segments supporting cyber ISR, OOB, and BDA in a closed, safe environment that would allow cyber warfighters to get the best possible active learning environment.

References

Blasch, Erik, and Micheline Bélanger. "Agile battle management efficiency for command, control, communications, computers and intelligence (C4I)." In *Signal Processing, Sensor/Information Fusion, and Target Recognition XXV*, vol. 9842, pp. 248-258. SPIE, 2016.

Christensen, Gitte Højstrup. April 2017. Interdisciplinary Perspectives on Special Operations Forces. In *The 2016 Royal Danish Defence College (RDDC) Conference* (No. 2017). Royal Danish Defence College.

Couretas, J.M., 2022. Cyber ISR and Analysis. In *An Introduction to Cyber Analysis and Targeting* (pp. 91-118). Cham, Switzerland: Springer.

Diakoumakos, J., Chaskos, E., Kolokotronis, N. and Lepouras, G., 2021, July. Cyber-Range Federation and Cyber-Security Games: A Gamification Scoring Model. In *2021 IEEE International Conference on Cyber Security and Resilience (CSR)* (pp. 186-191). IEEE.

Douglas, O.C., 2020. Transforming DOD for Agile Multi-Domain Command and Control. *Joint Force Quarterly*, *97*, pp.83-90.

Fowler, M. and Highsmith, J., 2001. The Agile Manifesto. *Software Development*, *9*(8), pp. 28-35.

Freitas, F., Silva, F.J., Campilho, R.D.S.G., Pimentel, C. and Godina, R., 2020. Development of a suitable project management approach for projects with parallel planning and execution. *Procedia Manufacturing*, *51*, pp.1544-1550.

Gartzke, E. and Lindsay, J.R., 2017. Thermonuclear Cyberwar. *Journal of Cybersecurity*, *3*(1), pp.37-48.

Giachetti, R.E. and Van Bossuyt, D.L., 2022. Challenges of Adopting DevOps for the Combat Systems Development Environment. *Defense AR Journal*, *29*(1), pp.22-49.

Gratch, Jonathan and Marsella, Stacy, 2003. Fight the Way You Train: The Role and Limits of Emotions in Training for Combat. *Brown Journal World Affairs*, *10*, p.63.

Highsmith, J. and Cockburn, A., 2001. Agile Software Development: The business of Innovation. *Computer*, *34*(9), pp.120-127.

Huang, P.M., Knuth, A.A., Krueger, R.O. and Garrison-Darrin, M.A., 2012, May. Agile hardware and software systems engineering for critical military space applications. In *Sensors and Systems for Space Applications V* (Vol. 8385, pp. 104-112). SPIE.

Huskey, T.W., 2007. *Impacts and Consequences of Non-Standard COTS C4I System Acquisition Upon Associated Programs of Record*. Naval Postgraduate School, Monterey, CA.

Jacobs, J., Romanosky, S., Suciuo, O., Edwards, B. and Sarabi, A., 2023. Enhancing Vulnerability Prioritization: Data-Driven Exploit Predictions with Community-Driven Insights. *arXiv preprint arXiv:2302.14172*.

Jelo, M. and Helebrandt, P., 2022. Gamification of cyber ranges in cybersecurity education. In *2022 20th International Conference on Emerging eLearning Technologies and Applications (ICETA)* (pp. 280-285). IEEE.

Kiras, James D., 2017. "Risky Business": A Conceptual Inquiry of Special Operations and Risk. *Interdisciplinary Perspectives on*, p.142-164.

National Initiative for Cybersecurity Education (NICE). 2018, *Cyber Ranges*. Viewed December 13, 2022, https://www.nist.gov/system/files/documents/2018/02/13/cyber_ranges.pdf.

NavalX. 2022. Centers for Adaptive Warfighting: Agile Training for Warfighters. https://www.secnav.navy.mil/agility/Pages/caw.aspx.

Novak, H., Likarish, D. and Moore, E., 2013. Developing Cyber Competition Infrastructure Using the SCRUM Framework. In *Information Assurance and Security Education and Training* (pp. 20-31). Springer, Berlin, Heidelberg.

Pereira, J.C. and de FSM Russo, R., 2018. Design Thinking Integrated in Agile Software Development: A Systematic Literature Review. *Procedia computer science*, *138*, pp. 775-782.

Schafer, Joseph H., Morrell, Chris and Blaine, Ray, 2022. The IWAR Range+ 21 Years: Cyber Defense Education in 2022. *Military Cyber Affairs*, *5*(1), pp.1-13.

Scrum.org. n.d. What is Scrum. [https://www.scrum.org/learning-series/what-is-scrum].

Scrum.org. 2023. The Scrum Framework Poster. https://www.scrum.org/resources/scrum-framework-poster.

Sullivan, D.T., Colbert, E.J.M., Hoffman, B.E. and Kott, A., 2018. Best practices for designing and conducting cyber-physical-system war games. *Journal of Information Warfare*, *17*(3), pp. 92-105.

Sutherland, J., & Sutherland, J. J. (2014). *Scrum: The Art of Doing Twice the Work in Half the Time*. New York. Crown Business.

Tomar, A.K., 2017. *Study of Scrum Framework Usage in Non-Software Systems* (Doctoral dissertation, Massachusetts Institute of Technology).

U.S. Cyber Range. (n.d.) What is the U.S. Cyber Range? https://kb.uscyberrange.org/faq/us-what-is-cyber-range.html.

Vykopal, J., Vizváry, M., Oslejsek, R., Celeda, P. and Tovarnak, D., 2017, October. Lessons learned from complex hands-on defence exercises in a cyber range. In *2017 IEEE Frontiers in Education Conference (FIE)* (pp. 1-8). IEEE.

A New Interpretation of Integrated Deterrence: Physical and Virtual Strategies

Jim Q. Chen
U.S. Department of Defense National Defense University, Washington, DC. USA

jim.chen@ndu.edu

Abstract: The integrated deterrence strategy, backstopped by nuclear deterrent, calls for seamless collaboration in deterrence across warfighting domains, using all instruments of national power, and with allies and partners. Being a warfighting domain and being closely related to the information instrument of national power, the cyber domain should certainly be included, and cyber deterrence should play a significant role in the integrated deterrence strategy. Nevertheless, as cyber deterrence seems not to be as effective as it is expected at least currently, some scholars and practitioners doubt its mere existence, not mentioning the role that it can play in the integrated deterrence strategy. This paper argues that not having deterrence in cyberspace leaves a blank spot in the strategy since some critical functionality of deterrence in cyberspace cannot easily be replaced. By recognizing the unique strategic context of cyberspace, the paper maintains that deterrent effect can actually be achieved in unique ways in this space. To further explore the unique role that deterrence in cyberspace plays within the integrated deterrence strategy, this paper proposes a multi-level and multi-aspect architecture for integrated deterrence strategy. This novel architecture is able to cover varied levels of strategic environments both below and above the threshold of armed conflict. It is also able to correlate varied deterrent measures with varied strategic environments categorized via various aspects, such as diplomacy, information, military, economy, etc. This paper shows that the inclusion of deterrence in cyberspace can empower the strategy by making the strategy flexible enough in tackling various challenges. Eventually, the strategy can make its contribution in preventing war and maintaining peace.

Keywords: Strategies, integrated deterrence, deterrence in cyberspace, multi-level and multi-aspect architecture, strategic contexts

1. Introduction

Deterrence is used to prevent war and maintain stability. It is an important means in the military instrument of national power. In the Joint Publication 3-0, *Joint Operations*, it is defined as "the prevention of action by the existence of a credible threat of unacceptable counteraction and/belief that the cost of action outweighs the perceived benefits." (The Joint Staff, 2011).

Kahn (1960), Schelling (1960), and Schelling (1966) are the pioneers of the deterrence theory. Kahn's work lays the foundation for deterrence-by-denial strategy, while Schelling's work lays the foundation for deterrence-by-punishment strategy. These two strategies, focused on nuclear deterrence, have shaped the current strategies regarding strategic stability. Payne (2008) provides a good explanation of these two strategies. In the cyber era, cyberspace has been created based on the physical sphere. One may certainly wonder whether deterrence can also be applied in cyberspace. If it does, to what degree it can be applied. If it does not, what other measures can be used in cyberspace.

Since it is very hard to create a virtual impact that can reach the level of nuclear deterrence, some scholars and practitioners question the mere existence of deterrence in cyberspace, not mentioning the role that it can play in the integrated deterrence strategy. One example is Harknett and Smeets (2020), in which they discuss the reasons for the failure of cyber deterrence as well as other measures that can be used in supporting some functionalities needed for cyber operations. Smeets (2020) further discusses the use of the replacement strategies such as persistent engagement and defend forward. Nevertheless, other scholars and practitioners insist on the pursuit of deterrence in cyberspace, as the crucial role of deterrence is not replaceable in their view. Healey (2019) points out the challenges in defend forward. Libicki (2009), Nye (2017), Chen (2017), Chen (2018a), Libicki (2021), and others argue for deterrence in cyberspace as well as its role in supporting national security. For instance, Libicki (2021; 2009) argues for cyber deterrence. He discusses the difference among cyber deterrence, nuclear deterrence, and criminal deterrence.

It needs to be pointed out that both schools of thoughts have some valid points in their arguments. Nonetheless, a solution that can be accepted by both sides is hardly to be reached, especially when the discussion is only focused on cyberspace.

To address this challenge, it is worthwhile to have a better understanding of the goal and the essential components of deterrence. The goal of deterrence is to force deterrent targets to give up their current aggressive commitment or at least restrain their behavior. In other words, it is to change the decision-making

calculus of deterrent targets. This is likely to happen when the following conditions are met: (1) The deterrent targets see no significant gains but only high costs in their commitment. (2) They are overwhelmed with fear and anxiety after they see no chance of changing the upcoming results and after they realize that the unavoidable consequences will threaten their survival and/or reputation. Art and Greenhill (2018) describe how coercion can lead to the change of an opponent's behavior. Brodie (1946) holds that atomic weapons are not made for use but to keep the opponents from using them. It is in this sense that it can be argued that deterrence works in the human minds.

One may wonder how such a psychological state can be created. To do so, the following essential components should be called upon. They are capabilities, willingness to use capabilities, demonstration of capabilities, and credible consequences. Successful deterrence relies entirely upon these components. Among these components, the component of capabilities is the most critical one. Without it, the other three components cannot play any role in deterrence.

One may then wonder what capabilities can be used to make deterrence work. For a long time, the strongest physical forces with technological superiority serve as the capabilities that create such a psychological state, because the potential physical punishment can fully convince deterrent targets that they have no chance of achieving their goal and their fate is doomed if they do not stop their commitment. In the physical world, nuclear weapons serve as a last resort in deterrence. This is because they, as weapons of mass destruction, can cause tremendous casualty and destruction to adversaries. The intimidating consequences of being attacked via these weapons are totally beyond what adversaries can bear in any circumstances. The destruction is absolute. The use of these weapons can have a destructive impact. When both sides possess these weapons and neither side ventures to use them first, a balance is thus created.

Currently, the notion of deterrence is heavily influenced by nuclear deterrence strategies, as nuclear weapons possess a paramount threatening capability. Nevertheless, the notion of deterrence itself does not exclude other ways of generating deterrent effects, at least theoretically. If there is a capability that can push deterrent targets into an extreme psychological state to stop or change their behavior, this capability can certainly serve as a capability for deterrence. It does not matter whether this capability is physical, digital, or the combination of the two. If it can help to achieve the deterrence outcome, it is worth to be explored.

To develop a novel solution in this paper, other aspects and instruments of national power should be considered. Extending the scope will generate insights that otherwise may not have. This paper is structured as follows: Here in Section 1, an overview of both deterrence and the major approaches in deterrence theories is provided. In Section 2, the issues of the current deterrence approach, especially the one in cyberspace, are examined. In Section 3, an innovative multi-level and multi-aspect architecture for the integrated deterrence strategy is proposed to address the issues discussed in the previous section. This architecture, incorporating deterrence in cyberspace, points out a new way of effectively implementing the integrated deterrence strategy and ultimately achieving the goal of the strategy. It contains the levels of severity both below and above the threshold of armed conflict as well as the aspects of varied instruments of national power, such as diplomacy, information, military, economy, law enforcement, allies, and partnership. In Section 4, the benefits of the multi-level and multi-aspect architecture for the integrated deterrence strategy are discussed. In Section 5, a conclusion is drawn.

2. Issues of Implementing Deterrence in Cyberspace

It should be pointed out that deterrent capabilities will be effective only when they are employed in the contexts they are designed for. Hence, it is critical to understand the relationship between a deterrent strategy and a strategic context, since one capability that works in one context is not as effective as it should be in another context.

Deterrence in the physical world possesses the following characteristics:

- The purpose of deterrence strategies is to prevent further escalation of conflict and avoid war.
- The deterrent capabilities designed, developed, and employed are for armed conflict. Hence, they are used above the threshold of armed conflict.
- The display of capabilities and the demonstration of willingness to use the capabilities are transparent and unambiguous.
- The entry level is high because tremendous investment is needed for the education of the professionals as well as the design and development of weapons with technological superiority.

- The main deterrent sources are physical, the potential destructive impacts indicated are physical and psychological, and the actual deterrent consequences are psychological and physical.

However, this is not exactly the same in cyberspace. As Chen (2019) points out, cyberspace possesses the following critical factors: speed, accuracy, precision, dynamics, stealth, ambiguity, and individual operators. In most cases, capabilities used for a deterrence purpose in cyberspace are seldom made known publicly. The impact is felt. The capabilities as well as the willingness to use capabilities are demonstrated, but the source codes and algorithms are always kept secret. Besides, as pointed out by General Paul Nakasone, USA, in Nakasone (2019a), "the locus of struggle for power has shifted towards cyberspace, and from open conflict to competitions below the level of armed attack". In addition, the entry level in cyberspace is low (Nakasone, 2019b).

Deterrence in the virtual world possesses the following characteristics:

- The purpose of deterrence strategies is to stop aggressive cyber operations from deterrent targets and prevent cyber conflict from further escalating into armed conflict.
- The deterrent capabilities designed, developed, and employed are for both non-armed and armed conflict. Hence, they are used both below and above the threshold of armed conflict.
- The display of capabilities is opaque and ambiguous, while the demonstration of willingness to use the capabilities is transparent and unambiguous.
- The entry level is relatively low because the cost of design and development of cyber weapons with technological superiority is relatively low compared with the cost of design and development of nuclear weapons (even though tremendous investment is needed for the education of professionals).
- The main deterrent sources are virtual; the potential destructive impacts indicated are virtual, physical, and psychological; and the actual deterrent consequences are psychological, virtual, and physical.

Comparing the characteristics in the virtual world with those in the physical world, one can quickly see that there are differences in the purpose of deterrence, the requirement for deterrent capabilities, the requirement for the display of capabilities, the requirement for the demonstration of willingness to use the capabilities, the requirement for the investment of human resources as well as research and development, deterrent resources, the potential destruction impacts indicated, and the actual deterrent consequences.

Given these differences, it is hard to imagine that the deterrence strategies designed solely for the physical world could be effective in cyberspace. The strategy of deterrence by denial and the strategy of deterrence by punishment are two examples.

The strategy of deterrence by punishment is hard to be employed in cyberspace because proportional punishment, immediate responses, and jurisdiction are usual challenges. In almost all cases, the chances of using nuclear responses to cyber breach operations are slim, since they are not at the same level of severity. Such responses are not proportional. In some cases, economic sanctions and diplomatic protests are introduced, but the responses are frequently delayed. In other cases, retaliatory cyber operations are launched. Due to the anonymous and hidden nature of some cyber operations and the amount of time needed for attribution, immediate responses are hard to be guaranteed.

The strategy of deterrence by denial requires strong defense in cyberspace. However, due to the complexity of systems (namely, having hidden layers that users do not have access to and/or having codes hard to be comprehended by users without technical knowledge) and the opaque nature of cyber supply chain, vulnerabilities within systems have been constantly exploited, thus defeating layers of defense. Restricted by the amount of time needed for attribution and accurate targeting, prompt retaliation in cyberspace is either delayed or is left aside. As a result, this deterrence strategy does not work in some cases.

Kello (2022) examines the two approaches in which deterrence is not sufficiently used in cyberspace as well as the consequences that these approaches may bring about. In one approach, nothing is done after being attacked in cyberspace. Kello (2022) calls this approach "appeasement", and considers it being "inherently escalatory". According to him, "nations that pursue it again teach opponents the lesson of 1938: further gains from aggression lie in wait". In the other approach, persistent engagement is employed while deterrence in cyberspace is not used. This approach "has a proactive element: it takes the fight to the other side's computer terrain". However, it is mainly for defensive purpose, not "a policy of attacking forward". Besides, it "commits unknown national resources to wherever the opponent brings the fight or is preparing it". Hence, following what is discussed in Kello (2017), Kello (2022) argues for punctuated deterrence, which punishes "a series of actions

and their cumulative effects". In Kello's view, punctuated deterrence "gives the victims the option to retaliate wherever and whenever fits their interests and capacities".

Kello's approach shows the significance of having deterrence in cyberspace as well as the difference between deterrence in cyberspace and persistent engagement in cyberspace. Nevertheless, this approach is primarily used for dealing with challenges in cyberspace. Its position in integrated deterrence is not investigated. Hence, this needs to be explored in the next section of this paper.

As discussed, deterrence in cyberspace is undoubtedly needed. However, in cyberspace, the strategy of deterrence by punishment is hard to execute, while the strategy of deterrence by denial may not be successful in all cases because these deterrence strategies are specifically designed for conflict in the physical world, and they are not specifically designed for cyberspace. Hence, they may or may not be effective in the cyber contexts.

As shown above, deterrent capabilities are effective only when they are employed in the strategic contexts they are designed for. So are the deterrence strategies that employ deterrent capabilities.

It needs to be pointed out that having problems in implementing these deterrence strategies in cyberspace does not mean that deterrence will never work in cyberspace. It instead means that appropriate deterrence strategies have not been employed in cyberspace as they have not been figured out and their corresponding deterrent measures have not been built yet. One of the urgent tasks that we have is to design context-based and appropriate deterrence strategies for cyberspace, building corresponding deterrent measures, and relating these strategies to the overall integrated deterrence strategy. This is what is going to be explored in the next section.

3. An Innovative Proposal

Varied deterrent capabilities not only are bound by varied contexts designed for but also can be tied together to achieve deterrence effects. When they are organized in a holistic way, a new version of integrated deterrence is brought into being.

In the 2022 U.S. National Security Strategy (NSS), integrated deterrence is defined as "the seamless combination of capabilities to convince potential adversaries that the costs of their hostile activities outweigh their benefits". It entails integration across domains, regions, the spectrum of conflict, the U.S. Government, and with allies and partners. In the 2022 U.S. National Defense Strategy (NDS) Fact Sheet, three primary ways of advancing the Department of Defense goals are mentioned. They are integrated deterrence, campaigning, and building enduring advantages.

In these national strategies, clearly listed are the key components of integrated deterrence, namely, the role of nuclear deterrent, cross-domain and cross-aspect interagency efforts, and cooperation with alliances and partners. In a sense, the integrated deterrence strategy, backstopped by nuclear deterrent, mobilizes all the parties and resources available, both internally and externally, to build enduring advantages in deterring adversaries.

Starling, Wetzel, and Trotti (2021) describe the integrated deterrence concept as an expansion from traditional deterrence to strategic deterrence by promoting whole-of-government deterrence plus whole-of-alliance deterrence. Within whole-of-government deterrence, instruments of national power are used.

To capture the essence of these strategies, this paper, following the model of deterrence levels recommended in Chen (2018a), proposes a multi-level and multi-aspect architecture, which is shown in Figure 1 below. This architecture consists of the levels that capture the escalation and de-escalation of conflict as well as the width that covers multiple instruments of national power and alliances.

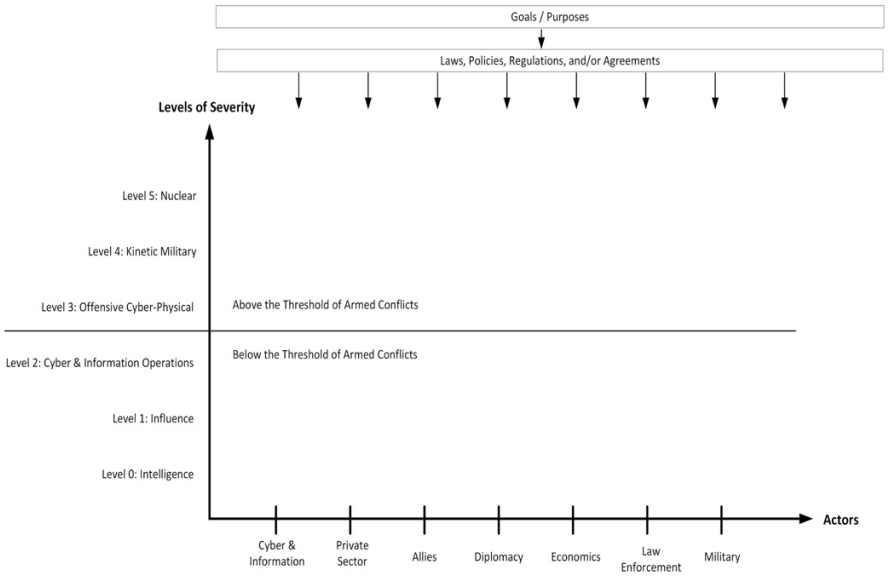


Figure 1: The Multi-level and Multi-Aspect Architecture for Integrated Deterrence Strategy

This architecture captures various levels of competition and conflict based on severity. These levels are either below or above the threshold of armed conflict. The following levels are below the threshold of armed conflict:

- Level 0 is the level of intelligence collection and analysis operations.
- Level 1 is the level of influence campaigns.
- Level 2 is the level of cyber operations and cyber-enabled information operations.

The following levels are above the threshold of armed conflict:

- Level 3 is the level of offensive cyber-physical operations and campaigns.
- Level 4 is the level of kinetic and conventional military campaigns.
- Level 5 is the level of nuclear warfare.

Going from a lower level to a higher lever indicates an escalation of a conflict and an increase in deterrence intensity, while going from a higher level to a lower level designates a de-escalation of a conflict and a decrease in deterrence intensity.

This architecture also captures the varied aspects that can be involved in deterrence endeavors. They consist of the following aspects: information, private sector, allies, diplomacy, economics, law enforcement, and military. Each aspect has a focus on a certain level or certain levels. Only the military aspect has the access to all levels. Besides having more aspects involved in the execution of deterrence means widening the scope of deterrence, while having fewer aspects involved in the execution of deterrence means narrowing the scope of deterrence.

Being dynamic and flexible, this architecture makes it possible to have different levels of deterrence in different strategic contexts. This means that there is no need to execute nuclear deterrence at Level 2, namely, the level of cyber operations and cyber-enabled information operations. However, as the situation escalates, the level of severity may get up to Level 5, namely, the level of nuclear warfare. Nuclear deterrence can be proportionally executed in an appropriate context. It is in this way that the nuclear deterrent is positioned in the integrated deterrence strategy, namely, supporting all lower level deterrent measures and serving as the ultimate deterrent measure. This satisfies the requirement of the 2022 NSS, which states that “nuclear deterrence remains a top priority for the Nation and foundational to integrated deterrence.”

Being dynamic and flexible, this architecture also makes it possible to widen the scope of deterrence. Instead of only resorting to the military force of one country, allied forces may be asked to get involved. By increasing the number of aspects involved, the weight of deterrence becomes amplified. Similarly, by engaging the private sector at the levels below the threshold of armed conflict, more capabilities can be assembled in executing deterrence in cyberspace.

In this architecture, Level 5 is focused on total physical destruction, while Level 2 is focused on cyber operations, especially surprise and signalling from the virtual world.

To summarize, this multi-level and multi-aspect architecture for the integrated deterrence strategy can satisfy the requirements of the 2022 NSS and the 2022 NDS by successfully bringing varied deterrent capabilities and varied strategic contexts together in a holistic way, thus achieving proportional and effective deterrence effects.

Given this architecture, deterrence must be proportional and effective at every level. Should there be little or no deterrent capabilities at one level, the integrated deterrence strategy will not be effective overall. Therefore, deterrence in cyberspace is required to support the overall integrated deterrence. Should cyber conflict at Level 2 escalate to cyber-physical conflict at Level 3, human casualty and property damage can be expected via attacks against critical infrastructure.

In Fischerkeller and Harknett (2018), two strategic spaces are mentioned. One is the strategic competitive space short of armed conflict, and the other is the strategic space of militarized crises and armed conflict. The first space contains a competitive interaction dynamic. It is in this space where the persistent engagement strategy and the defend forward strategy are executed. The second space encompasses escalation dynamics. It is in this space where compellence measures, including conventional weapons and nuclear weapons, are used. Deterrence lies in the space between these spaces. However, little is talked about the deterrence space. This approach is thus limited. Leaving a blank spot on the continuum is not a good idea, because one line of defense is left unused. Hence, ways of creating deterrence in cyberspace should be included for strategic reasons. The multi-level and multi-aspect architecture for integrated deterrence strategy elegantly bridges this gap, as it, sitting in the deterrence space, provides a link between the strategic competitive space short of armed conflict and the strategic space of militarized crises and armed conflict. With the introduction of this continuum, the role of deterrence prior to armed conflict can be explicitly captured and efficiently emphasized.

Within the deterrence space, one level of deterrent is simply not enough since the deterrent will be either too strong or too weak in a non-associated strategic context. For instance, if the nuclear deterrent is chosen as an option for deterrence in cyberspace, it will not be executed as a response for cyber breach attacks in almost all cases because it is too strong and not proportional. Likewise, if kinetic military campaigns are chosen as an option for deterrence in a nuclear war, it will not work because it is too weak and not proportional. Hence, it is important to employ context-based deterrent in a specific context.

Fischerkeller, Goldman, and Harknett (2022) also mention four important variables in the cyber strategic environment. These variables are “accessibility, availability, speed, and affordability.” They hold that persistent engagement, which is “able to effectively anticipate and persistently set the conditions of security” in one’s favour “in and through cyberspace”, is the key for success in cyberspace since it perfectly fits in the cyber strategic environment, which is “the product of interconnectedness, constant contact, and inherently reconfigurable terrain and capacity to act across and through that domain”. In other words, the persistent engagement strategy should be executed in the strategic competitive space short of armed conflict. Even though their analysis is not for deterrence in cyberspace, it is still applicable regarding the strategic context in cyberspace.

Which deterrent then fits in the cyber strategic context? Strategic surprise is one of them. This is because it can make deterrent targets overwhelmed with shock, confusion, and fear, at least for a period, and it can help to achieve the goal of stopping aggressive cyber operations from deterrent targets and preventing cyber conflict from further escalating into armed conflict. Strategic surprise can affect the minds of deterrent targets, making them lose the will to continue their commitment. This certainly offers deterrent initiators superiority during that period.

Kam (1988) points out that “surprise is a basic and recurring event in human life. Still, neither the repeated occurrence of surprises nor our assumption that life has surprises in store for us makes us any less vulnerable to its impact.” Following Kam (1988), strategic surprise can be created in the minds of adversaries when an act or development goes against the expectations of adversaries; catches them unprepared; and provokes shock, confusion, and fear in their minds. To satisfy these conditions, the state of uncertainty is needed. This state of mind, in turn, relies on ambiguity, anonymity, speed, and unique capabilities such as intelligence collection and analysis, precision, accuracy, and stealth maneuver. Chen (2017) and Chen (2018b) provide some detailed discussion about these capabilities. These are also the critical factors in cyberspace. Hence, in the virtual world, they can be naturally employed to create deterrent effects.

Therefore, in the strategic context of cyberspace, strategic surprise supported by stealth operations can be employed as a deterrent. This is one of the components in the strategy of deterrence by engagement and surprise, proposed in Chen (2017). For instance, this deterrent can perform the following functions to

overwhelm deterrent targets with fear and anxiety. First, various types of surprised warming messages can be sent to deterrent targets via unexpected means, to unexpected devices, and at unexpected times, indirectly indicating that the deterrent targets are being closely monitored and their identities are known. The frequency, target locations, message delivery methods, and message transmission time can be changed dynamically. Second, various direct and indirect but surprise cyber operations can be launched specifically against deterrent targets on varied hardware, software, and firmware without prior notice. Third, cyber-enabled information operations can be launched against deterrent targets without warning via unexpected means, to unexpected devices, and at unexpected times. Such operations are launched to damage deterrent targets' reputation via social media. All these surprised operations, which catch deterrent targets unprepared, can create uncertainty that will provoke shock, confusion, and fear. Overwhelmed by fear supported by uncertainty, deterrent targets can be deterred. Besides, varies operations send out varied signals to deterrent targets. Note that all these occur at Level 2 of the architecture proposed. These operations are below the threshold of armed conflict. It should be pointed out that the deterrence at this level is not as powerful as the deterrence at the level of nuclear weapons; however, it is proportional and serves its purpose. Besides, the signalling function is provided.

Deterrence at the offensive cyber-physical level is above the threshold of armed conflict. Here, cyber means are used for military purpose. Critical infrastructure is held hostage in deterrence. Nevertheless, the state of uncertainty can be generated with the help of ambiguity, anonymity, speed, and unique capabilities. This type of kinetic military operations or campaigns can also overwhelm deterrent targets with shock, confusion, and fear. Consequently, deterrent effect is created, and deterrent targets can be deterred. Likewise, extremely strong signals are sent out to deterrent targets.

As shown above, deterrent capabilities in cyberspace can be used either below or above the threshold of armed conflict. They are applicable not only in cyberspace but also in the physical world. With the multi-level and multi-aspect architecture for integrated deterrence, the differences in deterrent purposes, capabilities, and applications can be explicitly and successfully captured.

4. Benefits of the Multi-level and Multi-Aspect Architecture for the Integrated Deterrence Strategy

The multi-level and multi-aspect architecture for the integrated deterrence strategy holistically captures the essence of the 2022 NDS. It successfully reveals the relationship among different components, especially the dependency relationship. This makes it possible to execute deterrence proportionally and effectively within a certain strategic context, increasing its chance of being successful. This holistic approach also makes deterrence dynamic and flexible by resorting to varied deterrents based on varied strategic contexts. This architecture for integrated deterrence differentiates varied strategic contexts either below or above the threshold of armed conflict. Recognition of different strategic contexts helps to design and develop context-based deterrent capabilities, as well as dynamic tactics, techniques, procedures (TTP), and mechanisms. This approach makes deterrent capabilities efficient and effective within their relevant strategic contexts.

With the introduction of the multi-level and multi-aspect architecture for the integrated deterrence strategy, some challenging questions can be successfully addressed. In this new approach, a deterrent is no longer treated as something that can be executed anywhere. Instead, it must be executed within a relevant strategic context should it be effective. In this sense, they are context-based deterrent capabilities.

This holistic approach not only makes deterrence in cyberspace a part of the overall integrated deterrence but also provides national security decision-makers and strategists with ways of executing proportional and effective deterrence at different levels and within different aspects/scopes, thus successfully achieving national security goals. Ultimately, the research shows that deterrence in cyberspace is an essential component of integrated deterrence.

This innovative approach can certainly enrich integrated deterrence, since more domains, dimensions, and facets are added into this strategy, thus making the strategy more powerful and more flexible.

5. Conclusion

This paper examines the issues of the current deterrence approaches, especially the one in cyberspace. It argues that deterrent capabilities will be effective only when they are employed in the contexts in which they are designed for. Based on this argument and the requirements of the 2022 NSS and the 2022 NDS, an innovative multi-level and multi-aspect architecture for the integrated deterrence strategy is proposed. The need for

deterrence in cyberspace is discussed. So are various ways of conducting deterrence in cyberspace, especially ways that resort to strategic surprise. Besides, the benefits of the multi-level and multi-aspect architecture for the integrated deterrence strategy are discussed.

Given the proposed multi-level and multi-aspect architecture for integrated deterrence, varied deterrent capabilities and varied strategic contexts can be holistically tied together in a structured way; deterrence in cyberspace can be accounted for as an essential component in integrated deterrence; varied novel deterrent capabilities at different levels and within different aspects can be created, and ways of effectively using all deterrent capabilities at different levels can be provided to national security decision-makers and strategists in helping them to execute proportional and effective deterrence at various levels and within various aspects based on the requirements of strategic contexts, thus maintaining strategic advantage. As a result, the integrated deterrence strategy can eventually prevent war and maintain peace.

References

Art, R. and Greenhill, K. (2018) "Coercion: An Analytical Overview", *Coercion: The Power to Hurt in International Politics*, K. Greenhill & P. Krause (eds.), pp. 3–32, Oxford University Press, Oxford, UK.

Brodie, B. (1946) (ed.) *The Absolute Weapon: Atomic Power and World Order*, Harcourt Brace, New York.

Chen, J. (2019) "AI-Based Deterrence in the Cyber Domain", *Proceedings of the 14th International Conference on Cyber Warfare and Security*, pp.38–45.

Chen, J. (2018a) "On Levels of Deterrence in the Cyber Domain", *Journal of Information Warfare*, Vol.17,.2, pp.32-41.

Chen, J. and Dinerman, A. (2018b) "Cyber Capabilities in Modern Warfare", *Cyber Security: Power and Technology*, M. Lehto and P. Neittaanmäki (eds.), pp. 21–30. Springer.

Chen, J. (2017) "Cyber Deterrence by Engagement and Surprise", *PRISM*, Vol.7, No.2, pp.101-107.

Fischerkeller, M., Goldman, E., and Harknett, R. (2022) *Cyber Persistence Theory: Redefining National Security in Cyberspace*, Oxford University Press, New York, New York.

Fischerkeller, M. and Harknett, R. (2018) "Persistent Engagement, Agreed Competition, Cyberspace Interaction Dynamics, and Escalation", Institute for Defense Analyses.

Harknett, R. and Smeets, M. (2020) "Cyber Campaigns and Strategic Outcomes: The Other Means", *Journal of Strategic Studies*, Spring 2020 Issue, pp.1–34.

Healey, J. (2019) "The Implications of Persistent (and Permanent) Engagement in Cyberspace, *Journal of Cybersecurity*, 2019, pp.1-15.

Kahn, H. (1960) *The Nature and Feasibility of War and Deterrence*, P-1888-RC, RAND Corporation, Santa Monica, California.

Kam, E. (1988) *Surprise Attack: The Victim's Perspective*, Harvard University Press, Cambridge, Massachusetts.

Kello, L. (2022) *Striking Back: The End of Peace in Cyberspace – And How to Restore It*, Yale University Press, New Haven, CT.

Kello, L. (2017) *The Virtual Weapon and International Order*, Yale University Press, New Haven, CT.

Libicki, M. (2021) *Cyberspace in Peace and War*, (2nd Edition), Naval Institute Press, Annapolis, Maryland.

Libicki, M. (2009) *Cyberdeterrence and Cyberwar*, Project Air Force. RAND Corporation, Santa Monica, California.

Nakasone, P. (2019a) "A Cyber Force for Persistent Operations", *Joint Force Quarterly*, Vol.92, 1st Quarter 2019, pp.10–14.

Nakasone, P. (2019b) "An Interview with Paul M. Nakasone", *Joint Force Quarterly*, Vol.92, 1st Quarter 2019, pp.4–9.

Nye, J. (2017) "Deterrence and Dissuasion in Cyberspace", *International Security*, Vol.41, No.3, pp.44-71.

Payne, K. (2008). *The Great American Gamble: Deterrence Theory and Practice from the Cold War to the Twenty-First Century*, National Institute Press.

Schelling, T. (1960) *The Strategy of Conflict*, Harvard University Press, Cambridge, Massachusetts.

Schelling, T. (1966) *Arms and Influence*, Yale University Press, New Haven, Connecticut.

Smeets, M. (2020) "U.S. Cyber Strategy of Persistent Engagement & Defend Forward: Implications for the Alliance and Intelligence Collection", *Intelligence and National Security*, Vol.35, No.3, pp.444-453.

Starling, C., Wetzel, T., and Trotti, C. (2021) "Seizing the Advantage: A Vision for the Next US National Defense Strategy", Atlantic Council Strategic Paper, Atlantic Council Scowcroft Center for Strategy and Security, Washington DC.

The Joint Staff. (2011) *Joint Publication 3-0: Joint Operations*, Washington DC.

The United States Office of the Secretary of Defense (OSD). (2022) *Fact Sheet: 2022 National Defense Strategy*, Washington DC.

The White House. (2022) *National Security Strategy*, Washington DC.

Influence Diagrams in Cyber Security: Conceptualization and Potential Applications

Sabarathinam Chockalingam[1] and Clara Maathuis[2]
[1]Institute for Energy Technology, Halden, Norway
[2]Open University of the Netherlands, Heerlen, The Netherlands

sabarathinam.chockalingam@ife.no
clara.maathuis@ou.nl

Abstract. Over the last years, cyber-attacks are increasing in organizations especially due to the use of emerging technologies and transformation in terms of how we work. Informed decision-making in cyber security is critical to prevent, detect, respond, and recover from cyber-attacks effectively and efficiently. In cyber security, Decision Support System (DSS) plays a crucial role especially in supporting security analysts, managers, and operators in making informed decisions. Artificial Intelligence (AI)-based techniques like Bayesian Networks, Decision Trees are used as an underlying approach in such DSSs. Furthermore, Influence Diagrams (IDs) possess the capability to support informed decision-making based on its existing applications in other domains like medical. However, the complete capability and potential of IDs are not utilised in cyber security especially in terms of its explainable nature for different stakeholders and existing applications in other domains. Therefore, this research tackles the following research question: "What are potential applications of Influence Diagrams (IDs) in cyber security?". We identified applications of IDs in different domains and then translated it to design potential applications for cyber security issues. In the future, this will help both researchers and practitioners to develop and implement IDs for cyber security-related problems, which in turn will enhance decision-making especially due to its explainable nature for different stakeholders.

Keywords: Cyber security, Decision making, Incident response, Influence diagrams, Graphical model.

1. Introduction

Using technologies such as Artificial Intelligence (AI)/Machine Learning (ML), Internet of Things (IoTs), robotics redefines and enhances the way in which we work and live (Olan *et al.*, 2022). However, such technologies present numerous opportunities and at the same time some threats to both organizations and individuals. One such threat is the cyber threat which refers to any situation or event that may have the potential to negatively affect an organization's operations, assets, or people by impacting Confidentiality, Integrity, and Availability (CIA) of an information system. Notably, due to the increased usage of technologies, there has been a rise in cyber-attacks in both the Information Technology (IT) and Operational Technology (OT) environment (Lallie *et al.*, 2021; Prajapati and Singh, 2022). Successful cyber-attacks, which are not addressed effectively in a timely manner, can have dire consequences ranging from financial impact to reputational loss (Leroy, 2022). Therefore, it is extremely important to be aware of such threats and know how to deal with it effectively through relevant prevention, detection, response, and/or recovery measures that in turn would also lead to responsible and sustainable use of technologies in organizations.

To that end, Decision Support System (DSS) in cyber security help security analysts, managers, and operators to make informed decisions especially on risk mitigation measures/response strategies (Rees *et al.*, 2011; Buzdugan, 2020). For instance, Chockalingam et al. developed a decision support that help operators in distinguishing attacks and technical failures for the *"incorrect sensor measurements*" problem in floodgates (Chockalingam *et al.*, 2021). In such DSSs, AI/ML approaches like Bayesian Networks (BNs), decision trees, neural networks are used an underlying approach (Li, 2018; Chockalingam *et al.*, 2021). Importantly, the field of cyber security continued to evolve during recent years especially due to the existing capabilities as well as recent developments within the field of AI. This mainly support the application of its approaches for building intelligent solutions in cyber security that would be able to predict data breaches (Wilde, 2016), detect cyber-attacks (Adepu and Mathur, 2016), respond to cyber-attacks (Chockalingam, 2021), and/or recover from cyber-attacks (Manasa and Kumar, 2022) effectively and efficiently.

However, although a vast plethora of AI/ML-based techniques dealing and capturing uncertainty are used in different cyber security applications, the use of Influence Diagrams (IDs) in this domain, is scarce. An ID is a simple visual representation of a decision-making process (Howard and Matheson, 2005) and also a generalization of BNs which consists of both qualitative and quantitative components. The essential foundation to On the other hand, IDs possess the capability to be useful for cyber security applications especially based on their existing applications in other domains like agriculture (Jensen and Jensen, 2013), medical (Owens, Shachter

and Nease Jr, 1997; Baio *et al.*, 2006), safety (Matviykiv, 2013; Wang, Huang and Zhang, 2013) in addition to its explainable nature that aid effective visualization of complex decision problems to different stakeholders (Weflen, MacKenzie and Rivero, 2022), which is an essential aspect for cyber security decision making. Therefore, this paper aims to fill this gap by addressing the following research question (RQ): "What are potential applications of Influence Diagrams (IDs) in cyber security?". The research objectives (ROs) are:

- **RO1.** To identify existing applications of IDs in different domains.
- **RO2.** To translate the identified applications to design potential applications of IDs for cyber security problems.

The rest of this paper is structured as follows: Section 2 provides an essential foundation to IDs followed by our research methodology in Section 4. Section 5 describes identified applications of IDs in different domains followed by the potential applications of IDs in cyber security. Finally, Section 6 highlights conclusions and future work directions.

2. Background – Influence Diagrams: Key Components and An Example

Together with decision trees, IDs are an effective visual representation for decision models. The same underlying mathematical concept and processes are graphically represented differently in IDs and decision trees (Owens, Shachter and Nease Jr, 1997). IDs are an extension to BNs with two additional type of nodes including decision and utility in addition to uncertainty nodes (or chance) nodes (Åström *et al.*, 2014). ID consists of three different types of nodes which include: (i) uncertainty (or chance) nodes represented as oval, (ii) decision nodes as rectangle, (iii) utility nodes represented as diamond. Decision nodes associate with actions that the decision maker has direct influence over, whereas chance nodes reflect events that are not under the decision maker's control (Lacave, Luque and Diez, 2007). Utility nodes represent the preferences of the decision maker. Utility nodes cannot be parents of chance or decision nodes.

An example ID is shown in Figure 1. The uncertainty nodes in the shown example are: *"Cyber Security Risk Level"*, and *"Cyber Security Awareness".* The decision node in the shown example is *"Cyber Security Training Needed or Not"* and the utility node is *"Value"*. Furthermore, in this example ID, we provided example values to the Conditional Probabilities Tables (CPTs) of the uncertainty nodes and example utilities for different policies in the utility node. This ID can be used by managers to determine an expected utility when they decide to provide cyber security training or not. For instance, in this example, the expected utility when they make a decision to provide cyber security training is 4800, whereas the expected utility when they make a decision to not provide cyber security training is 300. In this case, they can make a choice to provide cyber security training, which provides highest expected utility compared to the other choice and therefore optimal.

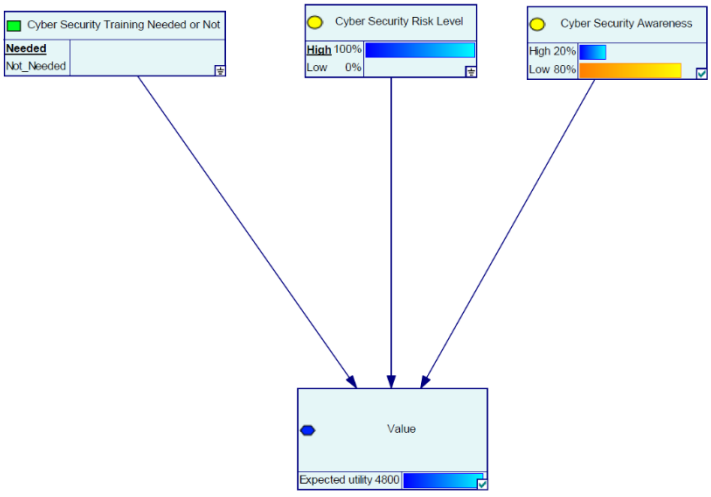


Figure 1. An Influence Diagram – Example

3. Research Methodology

The aim of this research is to provide awareness and understanding on the existing use of IDs in different application domains and potential applications of IDs in the cyber security domain. To that end, the following

RQ is formulated in this study: "What are potential applications of Influence Diagrams (IDs) in cyber security?". To address this RQ, an extensive literature review is conducted to build conceptual designs as a socio-technical artefact using the Design Science Research methodology (Peffers *et al.*, 2007; Peffers, Tuunanen and Niehaves, 2018). Accordingly, the following research phases are considered in this process.

In the first phase (i.e., problem definition and aim), the problem that should be investigated is analyzed in its corresponding context and the objective of this research is defined. Correspondingly, while a vast set of AI applications in the cyber security domain exists and their effectiveness was proved in tackling both knowledge and data characterized issues and challenges, still issues remain, old ones come back to live through new perspectives, and (un)expected new ones are born. However, a technique that is barely known and used in the cyber security domain are IDs (Li, 2018; Zhang *et al.*, 2022). To be able to investigate their potential applications in cyber security, an extensive literature review is conducted using keywords such as *"cyber security"*, *"influence diagrams"*, and *"application"* that are coupled in different combinations as queries used in scientific databases such as ACM, IEEE, Scopus, and Web of Science. Hence, an in-depth understanding of what IDs mean, how they can be characterized, and how they function is gathered followed by a series of various corresponding applications in different fields including agriculture, medical, military, and safety.

In the second phase (i.e., solution design), underlying structure with diverse types of nodes and model purpose are gathered from existing applications to build the artefact that this paper proposes, i.e., conceptual designs that captures the purpose, characteristics, and architecture of IDs for potential cyber security applications. In the third phase (i.e., evaluation), conceptual designs proposed is typically evaluated through instantiation/demonstration based on building a series in different cyber security applications such as cyber risk management. However, this phase is out of the scope of this paper. In the final phase (i.e., communication), the results in the context of this research in addition to possible future research directions are communicated by means of this publication and future presentations.

4. Existing Applications of Influence Diagrams

This section describes applications of IDs in different domains including agriculture, medical, military and safety.

Jensen et al. presented a Decision Support System (DSS) prototype based on IDs for the management of fungal disease (mildew) in winter wheat (Jensen and Jensen, 2013). In this prototype, there are four different variable types: (i) static information, (ii) dynamic information, (iii) decisions, and (iv) utilities. Static information includes variables like soil type, plant density, winter wheat variety, whereas dynamic information includes variables like weather, disease incidence, remaining time to harvest. Furthermore, decisions include dose of treatment variable and utilities include value of yield, cost of treatment, and value of disease induced yield loss. This prototype can help to determine the optimal decision on the dose of treatment considering evidence from other variables.

Baio et al. developed a decision model based on IDs for performing cost-effective analysis of influenza vaccination in the Italian elderly population (Baio *et al.*, 2006). In this decision model, strategy is the decision node which has the following states: do not vaccine, standard vaccine, and innovative vaccine. Furthermore, the decision node is influenced by other variables that are in the developed decision model like reduction in events generated by standard vaccine, reduction in events generated by innovative vaccine, occurrence rate of events, costs of GP visits, costs of standard vaccine, and costs of innovative vaccine. This decision model can help to determine the cost-effective decision on the influenza vaccination considering the evidence from other variables in the model. Furthermore, Owens et al. developed an ID for a medical decision problem on whether to perform PCR and whether to treat (Owens, Shachter and Nease Jr, 1997). In this ID, there are five different variables: (i) obtain PCR?, (ii) treat?, (ii) PCR result, (iv) HIV status, and (v) Quality Adjusted Life Expectancy (QALE). The first two variables correspond to decision nodes, whereas PCR result and HIV status correspond to deterministic node. Finally, variable is the utility/value node.

Wang et al. proposed a novel method based on IDs for fault troubleshooting of electromechanical products (Wang, Huang and Zhang, 2013). The ID for automotive engine fault troubleshooting proposed includes three different types of nodes: (i) uncertainty nodes, (ii) decision nodes, and (iii) utility nodes. Uncertainty nodes are fault causes of automotive engine failure which include variables like fire fault (ignition timing error, ignition signal cutting off, ignition coil failure). Decision nodes has the states repair and not repair. Utility nodes mainly provide information on the cost of the decision from the diagnostic engineers' perspective. This ID supports diagnostic engineers to repair a faulty component with highest expected utility to repair. Matviykiv et al. developed a DSS based on IDs for shock and vibration mitigation while drilling (Matviykiv, 2013). The developed

DSS based on ID consists of three diverse types of nodes: (i) uncertainty nodes, (ii) decision node, and (iii) utility nodes. Uncertainty nodes include variables such as actual life of a downhole tool, actual vibration severity, expected tool life at the end of the run after mitigation, expected vibration severity after mitigation. Decision node is whether to mitigate or not which has the following states: mitigate, ignore. Finally, there are two different utility nodes in the developed DSS based on IDs: utility function which reflects the level of tool preservation, utility function which reflects the negative effect of ineffective decision. Jeet et al. developed an approach based on ID for estimating staff training in software industry (Jeet *et al.*, 2009). The developed ID consists of four chance nodes: (i) newly appointed staff, (ii) lack of experience with project environment, (iii) lack of experience with project software, and (iv) staff not well versed with quality standards. In addition, the decision node is the staff training with the states "Yes" and "No". Finally, the utility node in the developed ID is the cost of conducting this training.

Tailby et al. discuss at a hypothetical level existing tensions between different nations (Tailby, Coyle and Gill, no date). They propose an ID, which integrates relevant political factors with military factors for military strategic planning through military capability development. To that end, they consider a scenario-based analysis for identifying critical missions and capabilities necessary to conduct missions robustly. This proposed model contains variables related to aspects such as defense preparation and Course of Action (CoA) execution, and through its extension integrates feedback loops. Moreover, this model is simulated by conducting war games in a multidisciplinary perspective, i.e., not only with military players, but also with diplomatic and government players. Based on the results obtained, this model proves to be: (i) useful for tackling different dimensions of uncertainty in such a complex domain like the military domain, (ii) helpful when presenting its results to its users, and (iii) supportive for generating conditions and assumptions for further analytical studies in this domain. Staker proposed a theoretical ID for military decision-making support for launching an attack on the enemy (Staker, 1999). This proposed model considers intelligence and enemy strength as uncertainty nodes (random variables), launch attack as the decision node, and commander satisfaction as the utility node. Bilusich et al. developed two IDs for conducting both military and civil-military operations based on the military adaptive campaigning strategy through gaining insights into multiple military and civilian related environments (Bilusich, Bowden and Gaidow, 2011). The proposed models make sure the coherent operational approach, development and communication of military Commander's intent in conducting the respective military operation/civil-military operation, distribution of roles, and the interactions between actors involved together with their definition. Among the variables embedded in the first model are law and order, lead humanitarian aid provision, and acceptable economic environment, and in the second model variables like military actors and population are included.

Bergdahl proposed an ID that allows integration of aircraft dynamics, preferences of the pilots, and the uncertainty of decision-making in a structural and transparent manner for providing military air combat support (Bergdahl, 2013). This model contains nodes such as threat situation assessment, combat state, and overall evaluation. With this model, its player (i.e., the pilot)) is able to analyse air combat tactics and manoeuvring that could further assist in autonomous decision-making processes embedded by systems like military air combat simulators. Mengmeng et al. developed an extended ID for evaluating Systems of Systems Architecture (SoSA) by using an anti-missile architecture case (Mengmeng *et al.*, 2018). This model contains type of nodes like the process phases of an anti-missile, radar characteristics such as receive guidance and detect threat, and battalion command like send command and receive preparedness. Moreover, taking into consideration the novelty of this proposed model and its probabilistic nature, further research is necessary for grasping an in-depth perspective on SoSA while case scenarios have to be further defined and simulated to obtain a proper and realistic probabilistic extraction from the case scenarios conducted. These above-mentioned ID applications in the military domain show that they can be used for different objectives that range from task assessment, insights gathering, to decision-making support for both the agents involved in conducting military operations as well as the ones impacted by their effects.

The relevant aspects from the abovementioned ID applications are adapted and used to suit the potential ID applications in cyber security which will be detailed in Section 5.

5. Potential Application of Influence Diagrams in Cyber Security

This section describes conceptual designs for different potential applications of IDs in cyber security that are translated from existing applications of IDs in other domains. We identified and described three different applications which include: (i) response selection, (ii) cyber security training, and (iii) risk management.

5.1 Application Design 1: Response Selection

In case an operator observes an undesired top event in CIs, an operator needs to determine the nature of cause (i.e., attack or fault) in addition to the root causes (i.e., attack vector in case of an attack and failure cause in case of a fault), and corresponding effective response strategies. For this purpose, we adapted the application from Wang et al. proposed for fault troubleshooting of automotive engine (Wang, Huang and Zhang, 2013) to suit our application as shown in Figure 2. In our application, the uncertainty nodes include attack and fault, which are the nature of the cause in addition to attack vectors and failure causes which are the root causes. This notion is mainly adapted to suit our application from Wang et al., where they used fault causes of automotive engine failure (example: ignition timing error, ignition signal cutting off) (Wang, Huang and Zhang, 2013). Furthermore, we also adapted decision node where they had the states repair and not repair. In our application, we chose decision nodes with states including different response alternates and no response. We used costs corresponding to each decision as the utility node. This ID can support operators to choose optimal response strategies corresponding to each attack vector/failure causes based on their expected utility. The key components of this application design includes: **Deterministic Node** – Undesired Top Event, **Uncertainty Nodes** – (i) Attack/Fault (Nature of the Cause); (ii) Attack Vectors$_{1...m}$ / Failure Causes$_{1...n}$ (Root Cause), **Decision Nodes** – Decision Node$_{A1..m}$ (Response alternates in addition to no response to each attack vector) / Decision Node$_{F1..n}$ (Response alternates in addition to no response to each failure cause). **Utility Nodes** – Cost (expected utility of each response decision).

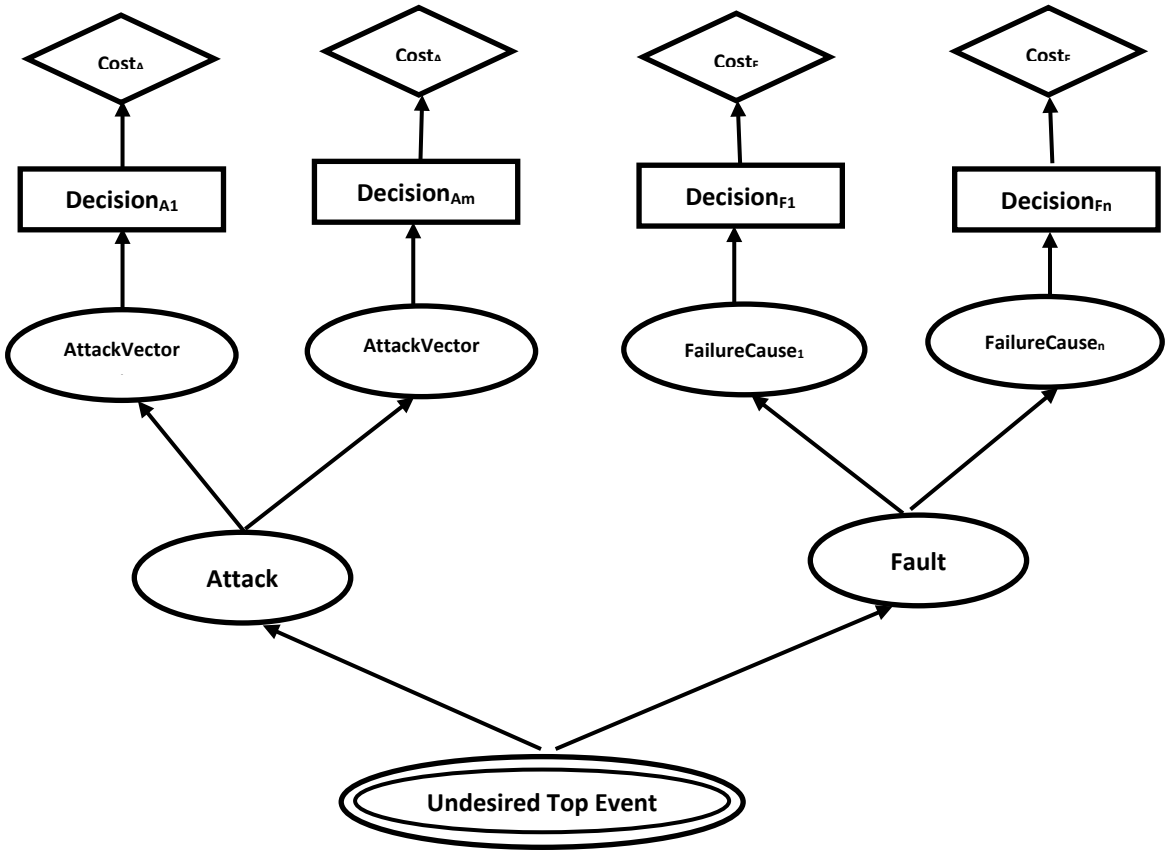


Figure 2. Design for Potential Application of IDs in Cyber Security – Response Selection

5.2 Application Design 2: Cyber Security Training

In organizations, it is important to provide decision support for management to prioritize group of personnel who need cyber security training and who do not need cyber security training. To that end, we adapted ID developed by Jeet et al. for estimating staff training in software industry (Jeet *et al.*, 2009) to suit our application as shown in Figure 3. They had four different chance nodes that are mainly related to skills and capabilities of a staff member (example: lack of experience with project software). We adapted it as skills and capabilities of a group of personnel in terms of cyber security in addition to the risk factors corresponding to a group of personnel and organization that suit our application. Furthermore, we adopted the notion of the decision node from (Jeet

et al., 2009) and used it as cyber security training with states needed or not needed. Finally, the decision node in our application corresponds to expected utility of providing cyber security training or not providing cyber security training. The key components of this application design include **Uncertainty Nodes** – competence level (within cyber security) (Beginner, Intermediate, Advanced); organization's cyber threat level (Low, Medium, High), handles confidential data (True/False), handles critical systems (True/False), **Decision Nodes** – cyber security training (True/False), **Utility Nodes** – expected utility of providing cyber security training and not providing cyber security training.

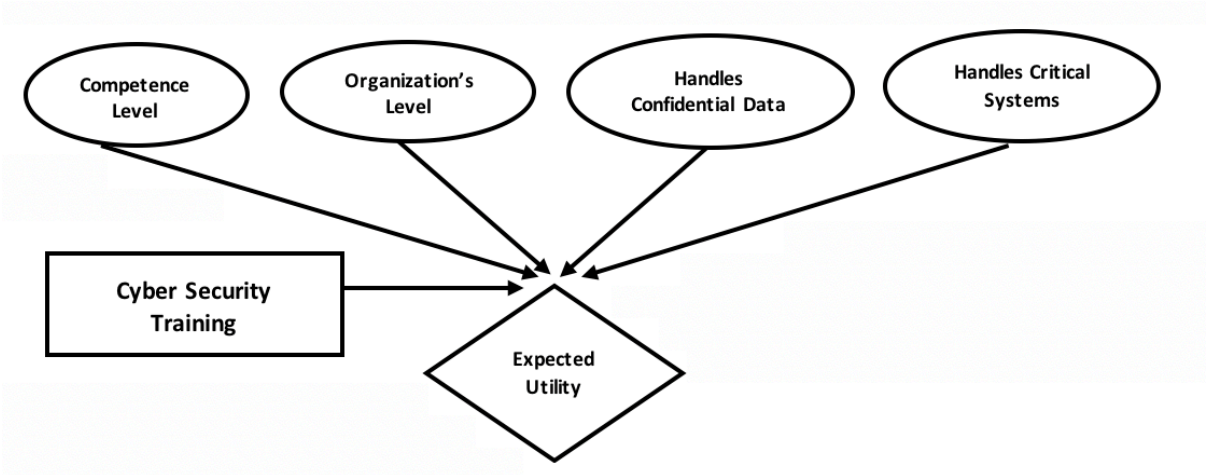


Figure 3. Design for Potential Application of IDs in Cyber Security – Cyber Security Training

5.3 Application Design 3: Cyber Risk Management

Cyber risk management plays an important role to put in place effective risk mitigation strategies. There is a need for decision support to put in place effective mitigation strategies to prevent a cyber-attack instead of responding to cyber-attack as in the case of our application design 1. The mitigation strategies could be to accept, avoid, transfer, or mitigate a risk depending on their impact and likelihood. We adapted the key components and structure from existing IDs proposed by Matviykiv et al. and Baio et al. (Baio *et al.*, 2006; Matviykiv, 2013) as shown in Figure 4.

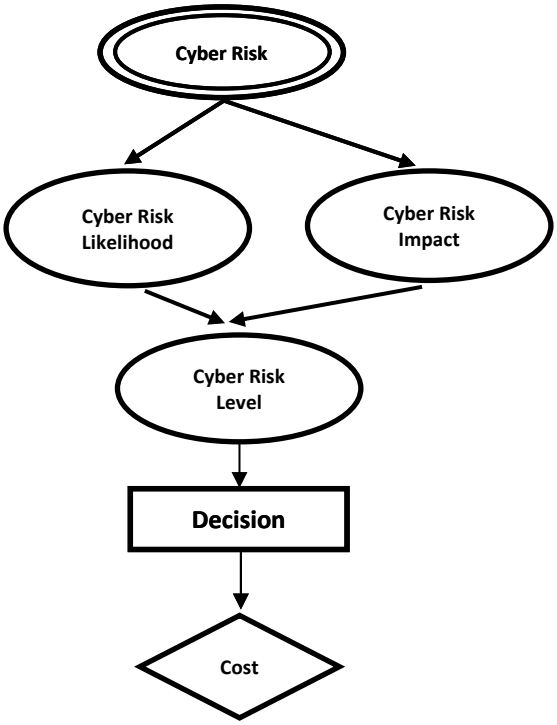


Figure 4. Design for Potential Application of IDs in Cyber Security – Cyber Security Training: Risk Management (Variant 1)

The key components of our variant 1 design include **Deterministic Node** – cyber risk (specific risk: true/false), **Uncertainty Nodes** – cyber risk likelihood (Likely/Possible/Unlikely), cyber risk impact (Significant/Moderate/Minor), cyber risk level (High/Medium/Low), **Decision Node** – Decision (Risk Mitigation Strategy: Accept/Avoid/Transfer/ Mitigate), **Utility Node** – Cost (this is the total cost associated with implementing risk mitigation strategy and ineffective decision).

In some cases, it is also important to choose a specific mitigation measure between different alternates within each mitigation strategy. We need to decide which mitigation measure to put in place in case we choose the strategy to mitigate the risk as our strategy. Therefore, we provide the design for this purpose as variant 2. The only difference between variant 1 and 2 is the states of risk mitigation strategy. Here in variant 2, we mainly look into specific risk mitigation measures instead of overall risk mitigation strategy. The key components of variant 2 include **Deterministic Node** – cyber risk (specific risk: true/false), **Uncertainty Nodes** – cyber risk likelihood (Likely/Possible/Unlikely), cyber risk impact (Significant/Moderate/Minor), cyber risk level (High/Medium/Low), **Decision Node** – Decision (Risk Mitigation Measure: different risk mitigation measures instead of high level strategy), **Utility Node** – Cost (this is the total cost associated with putting in place a specific risk mitigation measure and ineffective decision).

6. Conclusions and Future Work Directions

DSS provides an important foundation to make informed decisions within cyber security. Such DSSs are mainly developed using AI-based approaches like BNs, neural networks, and decision trees. IDs have the capability to be an effective underlying approach for DSS within cyber security especially due to its existing applications in other domains like medical in addition to its feature of easily explainable to different stakeholders ranging from personnel with technical and non-technical background. However, there is a lack of ID applications within cyber security. Therefore, in this research, we identified different existing applications of IDs in other domains including agriculture, medical, military, and safety. We then translated the identified existing applications of IDs into designs for potential applications in cyber security including response selection, cyber security training, and cyber risk management. In the future, we will implement provided designs and measure the effectiveness of IDs in cyber security applications through use-case approach.

References

Adepu, S. and Mathur, A. (2016) 'Using process invariants to detect cyber attacks on a water treatment system', in. *ICT Systems Security and Privacy Protection: 31st IFIP TC 11 International Conference, SEC 2016, Ghent, Belgium, May 30-June 1, 2016, Proceedings 31*, Springer, pp. 91–104.

Åström, H. *et al.* (2014) 'An influence diagram for urban flood risk assessment through pluvial flood hazards under non-stationary conditions', *Journal of water and climate change*, 5(3), pp. 276–286.

Baio, G. *et al.* (2006) 'Object-oriented influence diagram for cost–effectiveness analysis of influenza vaccination in the Italian elderly population', *Expert Review of Pharmacoeconomics & Outcomes Research*, 6(3), pp. 293–301.

Bergdahl, C. (2013) 'Modeling Air Combat with Influence Diagrams'.

Bilusich, D., Bowden, F.D. and Gaidow, S. (2011) *Applying influence diagrams to support collective C2 in multinational civil-military operations*. DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION EDINBURGH (AUSTRALIA).

Buzdugan, A. (2020) 'Review on use of decision support systems in cyber risk management for critical infrastructures', *Journal of Engineering Sciences*, (3), pp. 134–145.

Chockalingam, S. *et al.* (2021) 'Bayesian network model to distinguish between intentional attacks and accidental technical failures: a case study of floodgates', *Cybersecurity*, 4(1), pp. 1–19.

Chockalingam, S. (2021) 'Using Decision Trees to Select Effective Response Strategies in Industrial Control Systems'.

Howard, R.A. and Matheson, J.E. (2005) 'Influence diagrams', *Decision Analysis*, 2(3), pp. 127–143.

Jeet, K. *et al.* (2009) 'An Influence Diagram Based Approach for Estimating Staff Training in Software Industry', *Journal of Intelligent Systems*, 18(4), pp. 267–284.

Jensen, A.L. and Jensen, F.V. (2013) 'MIDAS-an influence diagram for management of mildew in winter wheat', *arXiv preprint arXiv:1302.3587* [Preprint].

Lacave, C., Luque, M. and Diez, F.J. (2007) 'Explanation of Bayesian networks and influence diagrams in Elvira', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(4), pp. 952–965.

Lallie, H.S. *et al.* (2021) 'Cyber security in the age of COVID-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic', *Computers & security*, 105, p. 102248.

Leroy, I. (2022) 'The relationship between cyber-attacks and dynamics of company stock: the role of reputation management', *International Journal of Electronic Security and Digital Forensics*, 14(4), pp. 309–317.

Li, J. (2018) 'Cyber security meets artificial intelligence: a survey', *Frontiers of Information Technology & Electronic Engineering*, 19(12), pp. 1462–1474.

Manasa, S. and Kumar, K.P. (2022) 'Digital Forensics Investigation for Attacks on Artificial Intelligence', *ECS Transactions*, 107(1), p. 19639.

Matviykiv, T.M. (2013) 'Use of influence diagrams for decision support in drilling automation', *Journal of Global Research in Computer Science*, 4(4), pp. 1–7.

Mengmeng, Z. *et al.* (2018) 'Functionality evaluation of system of systems architecture based on extended influence diagrams', *Journal of Systems Engineering and Electronics*, 29(3), pp. 510–518.

Olan, F. *et al.* (2022) 'Artificial intelligence and knowledge sharing: Contributing factors to organizational performance', *Journal of Business Research*, 145, pp. 605–615.

Owens, D.K., Shachter, R.D. and Nease Jr, R.F. (1997) 'Representation and analysis of medical decision problems with influence diagrams', *Medical Decision Making*, 17(3), pp. 241–262.

Peffers, K. *et al.* (2007) 'A design science research methodology for information systems research', *Journal of management information systems*, 24(3), pp. 45–77.

Peffers, K., Tuunanen, T. and Niehaves, B. (2018) 'Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research', *European Journal of Information Systems*, 27(2), pp. 129–139.

Prajapati, S. and Singh, A. (2022) 'Cyber-Attacks on internet of things (IoT) devices, attack vectors, and remedies: a position paper', *IoT and cloud computing for societal good*, pp. 277–295.

Rees, L.P. *et al.* (2011) 'Decision support for cybersecurity risk planning', *Decision Support Systems*, 51(3), pp. 493–505.

Staker, R. (1999) *Military information operations analysis using influence diagrams and coloured Petri Nets*. ELECTRONICS RESEARCH LAB SALISBURY (AUSTRALIA).

Tailby, M.D., Coyle, G. and Gill, A. (no date) 'The Application of Influence Diagrams for the Development of Military Experiments'.

Wang, Y.S., Huang, Y.P. and Zhang, R.J. (2013) 'Decision Strategy for Fault Troubleshooting Using Bayesian Influence Diagram', in. *Applied Mechanics and Materials*, Trans Tech Publ, pp. 541–545.

Weflen, E., MacKenzie, C.A. and Rivero, I.V. (2022) 'An influence diagram approach to automating lead time estimation in Agile Kanban project management', *Expert Systems with Applications*, 187, p. 115866.

Wilde, L. (2016) 'A Bayesian Network Model for predicting data breaches caused by insiders of a health care organization'.

Zhang, Z. *et al.* (2022) 'Artificial intelligence in cyber security: research advances, challenges, and opportunities', *Artificial Intelligence Review*, pp. 1–25.

Permission-Based Classification of Android Malware Applications Using Random Forest

Nikolaos Chrysikos[1], Panagiotis Karampelas[1] and Konstantinos Xylogiannopoulos[2]
[1] Hellenic Air Force Academy, Dekeleia, Greece
[2] University of Calgary, Calgary, Canada
nikoschrysikos2000@gmail.com
panagiotis.karampelas@hafa.haf.gr
kostasfx@yahoo.gr

Abstract: Android is arguably the most widely used mobile operating system in the world. Due to its widespread use, it has attracted a lot of attention of cybercriminals who attempt to exploit its architecture and outsmart innocent users to install malware applications. The number of such applications is growing every day either by alternating a basic exploitation mechanism or by creating novel mechanisms to exfiltrate users' data. As a result, there is an increasing need for detection mechanisms that can classify these applications to families based on their characteristics. A significant amount of research has already been devoted to analysing and mitigating this growing problem; however, this situation demands more efficient methods with higher precision. The paper proposes such a framework for analysing and classifying a malicious application to certain families relying on the permissions used. The proposed method involves the pre-processing of the applications to extract their permissions, the tokenization of permissions, the data cleansing and finally the application of the Random Forest Classifier to classify the applications in families. The proposed method is trained and tested with a dataset of 11,159 malicious applications categorized in 33 unique families. The precision, recall and f1-score achieved is 98%. The results of the proposed methodology are promising, since it even works in an unbalanced dataset and in many cases outperform other state-of-the-art approaches.

Keywords: Malware Classification, Android, Permissions, Machine Learning, Random Forest

1. Introduction

Presently, the prevalence of smart devices is universal constituting the most popular mean for Internet access worldwide. According to Statista (Petroc, 2023), 6.259 billion people owned a smart phone in 2021 which is translated to the 80% of the world population. Thus, people of diverse age groups with different educational backgrounds and origin have integrated smart phones in their everyday life for audio and video calling, instant messaging, carrying out financial transactions, etc. With this plethora of uses, smart phones become our personal data depositories in which all aspects of our lives are recorded. This fact has led cybercriminals to constantly target smart phones by attempting to infect them with malware in order to get hold of the personal data of the users for their personal gain. The 2021 Mobile Threat Report by Kaspersky Labs (Kaspersky, 2023) reports that more than 46 million attacks were carried out worldwide in 2021 by 3.5 million malicious mobile installation packages for Android. This operating system is installed in approximately the 70% of all the mobile devices worldwide in 2022 (Petroc, 2023) and thus, the majority of the malicious applications target Android devices.

Conversely, many security specialists identifying the activities of cybercriminals have proposed diverse defensive techniques in order to protect the mobile users and their personal data. One of the prominent techniques that many security researchers consider as more effective for protection is the automatic malware detection and classification since by recognizing the type of a malicious application, the security system can automatically take the necessary counter-measures. Usually, there are two approaches for malware classification either using dynamic or static analysis. In dynamic analysis, the behaviour of the application is monitored and specific characteristics are used as features in order to classify the application as malicious or benign while in static analysis, characteristics such as opcodes, permissions, receivers, etc. are extracted and checked. Both approaches have been proved adequately efficient to classify malicious applications but still they lack in accuracy (Ahmad et al, 2017). Moreover, various obfuscation techniques recently used by the attackers in the mobile application's code including encryption, compression, polymorphism and metamorphism have made their detection harder and thus novel detection techniques that overcome the specific countermeasures are needed.

Motivated by this gap, our approach is focused on the static analysis of permissions which is one of the most important characteristics of an Android application. The permissions are used to provide access to various resources and services that are available in the mobile phone (Amer, 2021). Usually, they are classified as normal, signature, special and dangerous. In the first category, permissions such as ACCESS_WIFI_STATE or INTERNET can be found which frequently are necessary for the operation of an application. In the second

category, permissions such as BATTERY_STATS or BIND_WALLPAPER are found which need to be signed with the same signature as the application that they ask to get access. The appearance of these permissions depends on the need for bindings with other applications or services. In the special category, permissions that require mobile user authorization are listed such as WRITE_SETTINGS. For these permissions, the users would have to provide their consent in order to be activated. As dangerous permissions are considered all those which enable the application to use potential private information or services in the mobile phone such as CAMERA, ACCESS_BACKGROUND_LOCATION, etc. Google lists 40 such permissions in the developers website[1] that they should be avoided unless it is necessary. In addition, the developers can define their own custom permissions to some users or other external applications. As it can be understood, the permissions' model is very complex and important for the security of the mobile application and that is why many researchers have focused their attempts for classifying malware applications using this characteristic (Adebayo and Aziz, 2015). In this line of research, our proposed methodology uses permissions to perform Android malware classification. Our methodology:

- uses only permissions as input data in order to analyse and classify the malware applications achieving at the same time very high accuracy. This is a competitive advantage since the vast majority of real world malicious Android applications have most of their data obfuscated, and thus an approach that would require more features could perform poorly or not at all if some of the data needed, are obfuscated or missing;
- can perform equally accurately in classifying applications independently of the homogeneity of the sample. In the experiment we run, we managed to classify accurately applications in families of more than 3,000 applications but also in families with less than 5 applications.

The paper is organized as follows: in the related work section, the contemporary research in malware classification and more specifically in permission-based malware classification is presented. In Section 3, the proposed methodology is explained and the corresponding analysis phases are described. Section 4 presents the experimental environment, the dataset and the computer specifications used for the experiment. The next section presents and discusses the results acquired presenting the key findings and the limitations of the proposed method. In the final section, the conclusions and tentative future work in this area are presented.

2. Related Work

Malware detection and classification as mentioned is one of the issues in mobile devices security that has attracted researchers attention the latest years. There are several research teams that have proposed different approaches in addressing the issue with diverse results. In our review of the related work, we approach the subject more broadly examining first some generic malware classification methods and then we delve into the those which use permissions.

2.1 Malware classification

There are diverse features of the mobile applications that can be used in order to classify malware applications. Bytecode, opcode, system API calls, permissions, etc. are typically used as classifiers in malware classification. In this section, some of the recent techniques utilizing diverse features are presented and discussed in comparison to the proposed methodology.

One approach that is frequently used in malware classification is that of detecting similarities in applications' bytecode. One such approach creates a bytecode based model, performing a frequency analysis in order to classify malicious applications to malware families (Kang et al, 2013). Eventually, machine learning algorithms such as Decision Trees, Association Rules Mining, Genetic Algorithms, etc. were used to categorize malicious applications scoring 94% of F-measure. If the Android applications are obfuscated, the specific methodology will provide poor results since each variant will have different bytecode due to the obfuscation. Another approach using system calls with call logs to classify mobile applications with Naïve Bayes classifier achieved a 99.86% accuracy applied in a sample of 5,560 applications (Ahmad et al, 2017). While the specific technique achieves very high accuracy, there are cases in which malware applications do not make phone calls and thus the specific approach may not be proved so efficient. In another approach, a sample of 13,850 mobile applications including

[1] https://developer.android.com/reference/android/Manifest.permission

3,960 malware and 1,108 benign application was used in a machine learning based technique in order to classify them as malware or benign using 36 features extracted by mobile security framework. The proposed technique achieved an accuracy of 93.63% (Sachdeva et al, 2018). However, the limitation of this approach is that the applications are not classified to their appropriate family but are detected as malware or not which in some cases is not adequate. A classification method using system calls to GPS data is used in another study on a sample of approximately 5,560 malware applications and 500 random applications in which the researchers found 32 related patterns that were used from the malware applications (Saudi, 2017). In a recent study, Arif et al (2021) used a sample of 10,000 Android applications in order to classify malware applications using static analysis. The feature used for the machine learning model was permissions and the Random Forest classification algorithm achieved accuracy 91.59% which is lower than the proposed methodology achieved using the same classification algorithm.

2.2 Permission-based classification

A first attempt to use permissions for malware description using diverse machine learning algorithms is described by Huang et al (2013). In the specific study, the researchers used 124,769 benign applications and 480 malicious applications. The achieved accuracy of the algorithms as reported is above 80%. Another similar system, called Permission-based Malware Detection System used permissions as behavioural markers and the researchers created a classifier taking into consideration the combination of application permissions to classify malicious applications. They tested their method in a relatively small dataset of 2,950 benign and malicious applications with an average precision of 92-94% (Rovelli & Vigfússon, 2014). Milosevic et al (2017) attempted to analyse and identify malicious applications for the Android operating system using a machine learning aided approach for static analysis that uses permissions. They have used a wide range of ensemble machine learning algorithms and the highest precision, recall and f1-score that they achieved was 89%. An ensemble machine learning approach for android malware classification using hybrid features was presented by Pektaş & Acarman (2018) utilizing static features such as permissions, hidden payloads and dynamic features such as API calls, network connections, etc. The authors applied a number of machine learning algorithms on a 3,339 malicious applications sample of 20 malware families achieving 92% accuracy. The whole analysis and detection process required 8 days. The SIGPID malware detection system is based on permission usage analysis that employees 22 significant permissions for malware classification (Li et al, 2018). The system is based on the Support Vector Machine classifier which achieves an accuracy 93.62% using these 22 permissions. The accuracy of the proposed methodology in this paper using a different classifier and all the relevant permissions achieves better accuracy.

3. Methodology

The proposed methodology implements a composite classification model that analyses Android applications, extracts data and eventually classifies malicious applications to pre-defined malware families. The malware detection approach is based on extracting and analysing application features from the AndroidManifest.xml file. We have divided our methodology into three phases. The initial phase starts with the reverse engineering of the applications and data extraction from their AndroidManifest.xml file. In the second phase, the acquired data are processed and transformed to the appropriate form in order to be fed to the machine learning algorithm for parsing in the third phase. Figure 1 illustrates the overall process of the proposed methodology.

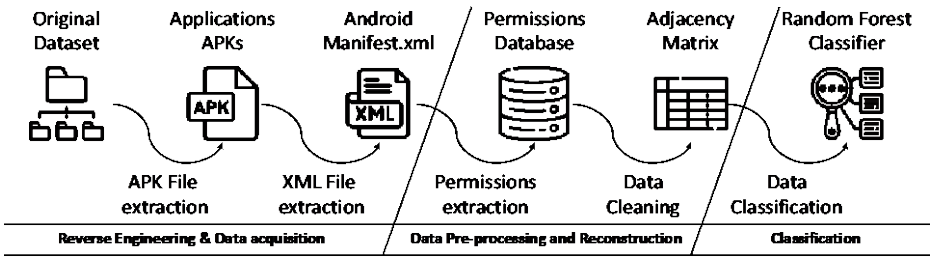


Figure 1: The overall process

3.1 Reverse Engineering and Data acquisition

The classification requires the training of the algorithm with a labelled dataset. For that purpose, we acquired a dataset of malicious Android applications (Wei et al, 2017) in order to perform the required training of the

adopted machine learning algorithm. Based on the description of the dataset, the applications are labelled and correctly classified to the corresponding malware families by the dataset creators. The training and test data ratio that was chosen was 75:25. From each family ,we removed the 25% of the applications in order to form the test dataset while the remaining apps were used to prepare the training dataset. However, in order to use the chosen features, further pre-processing is required.

The malicious applications were provided in the form of Android application PacKages (APKs) organized in families folders and thus, the AndroidManifest.xml file which contains the permissions was not immediately available. Therefore, reverse engineering was required to extract all the contents of each application package in order to acquire the necessary file. This was achieved by using Apktool[2] to extract the AndroidManifest.xml file. The second step required the collection of all the permissions of each malicious application from the manifest file. In this step, we used the list of the 206 Android permissions[3] but we also added any custom permission that was defined in the collected manifest files. In total, 358 unique permissions were collected.

3.2 Data Pre-processing and Reconstruction

In the second phase, we started the pre-processing of the data which includes the following activities:

- Duplicate Permissions Removal. We found that some malicious applications contained the same permission twice. We considered these occurrences as abnormal which will affect negatively the training of the machine learning algorithm and thus we removed them. The collected permissions were also sorted alphabetically for easier visual examination and comparison if required.
- Preparation of the Adjacency Matrix of Application and Permissions. The columns of the matrix represent the 358 permissions and each row of the matrix corresponds to each application. If a permission existed in the AndroidManifest.xml, this was denoted by 1 in the corresponding column of the permission or by 0 if it was not present. Instead of using the name of each individual application, we used as label the name of its malicious family. In Table 1 part of the adjacency matrix is demonstrated.
- Removal of unused permissions. Further data cleansing was required since there were permissions that were not used by any application and therefore they did not add any value to the next stage. Similarly, applications that do not request any permissions and consequently we would not be able to classify them were removed. More precisely, there was one such an application and 224 permissions that could influence negatively the next phase, so they were removed. Eventually, the final shape of the adjacency matrix was 11,160 by 134.
- Removal of insignificant features. In further preparation of the dataset for classification, we computed the hamming distance (Hamming, 1950) of features. If any pair of features has zero hamming distance, one of them was removed, since both columns conveyed exactly the same information. By this process 15 pairs were detected and in total 10 features were removed.

Table 1: Excerpt from the Adjacency Matrix of the features

Applications	GLOBAL_SEARCH	DEVICE_POWER	GET_PACKAGE_SIZE	MANAGE_USB
AndroRAT	0	0	0	0
AndroRAT	0	0	0	1
AndroRAT	1	0	0	0
AndroRAT	0	0	1	0

3.3 Classification

Random Forest Classifier (Breiman, 2001) has been used as the classification model of the methodology. The specific classifier is based on an “ensemble” of independent decision trees. Each tree is presented as a flow chart in which there are branches which represent the features and nodes that represent the different instances that

[2] https://ibotpeaches.github.io/Apktool/
[3] https://developer.android.com/reference/android/Manifest.permission

occur according to the result of the test (Panigrahi et al, 2018; Biau et al, 2008). In that sense, random forests (RF) combine these decision tree classifiers and based on the overall result of the decision trees it arrives in a final result. Random Forest was selected because it outperforms other similar machine learning classification techniques (Multi-layer Perceptron classifier, Balanced Bagging Classifier, Balanced Random Forest Classifier, Support Vector Machine, K-nearest neighbors classifier) (Table 2) since it works well with high volume data with many classes and low correlation, as these in the available dataset, and produce high precision in classification tasks (Liu et al, 2012). The RF Classifier that was used is implemented in Python as part of the Scikit-Learn[4]. To apply the specific classifier, the original dataset was split into two as mentioned in a ratio of 75:25.

Table 2: f1 – score comparison of Random Forest against other known classifiers

	KNN	SVM	RF	BRFC	BBC	MLP
accuracy	0.96	0.95	0.98	0.74	0.64	0.95
macro avg	0.84	0.69	0.97	0.57	0.50	0.70
weighted avg	0.95	0.94	0.98	0.77	0.64	0.95

4. Experiments

The dataset used originally consisted of 24.650 malware samples for Android operating system categorized in 71 families (Wei et al, 2017) which were collected between 2010 and 2016. From the 71 families, we used 33 families for training and testing since some of the rest contained corrupted data and it was not possible to use them. In addition to that, it should be noted that not all samples of the 33 families contained an AndroidManifest.xml file which means that the actual number of samples that we used were less than the number of applications that a family actually contained in the original dataset. Table 3 provides an overview of the malware families we used in our dataset. For each family, its corresponding type is displayed indicating its main purpose. In total, these 33 families contained 14,437 samples from which only the 10,960 malicious applications were used, which is the 75.9% of all the applications in the 33 families.

The computing environment used for the analysis consisted of a laptop with an AMD Ryzen™ 7 3750H mobile processor, a Radeon Vega Mobile Gfx graphics processor and 16Gbs of memory. The overall experiment lasted about 2 hours starting with the initial extraction of the permissions from the AndroidManifest.xml file of each application, pre-processing, classifying and repeating the process for all samples. On average, during the experiment the resources of the laptop used were approximately 6Gbs of memory and 70% of the computing power of the processor.

Table 3: The dataset statistics

Families	Type	Samples #	Used Samples #	Percentage
AndroRat	Backdoor	46	46	100
Andup	Adware	45	45	100
Aples	Ransom	21	21	100
BankBot	Trojan-Banker	740	647	87.3
Bankun	Trojan-Banker	72	72	100
Boqx	Trojan-Dropper	221	221	100
Boxer	Trojan	44	44	100
Cova	Trojan-SMS	18	18	100

[4] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

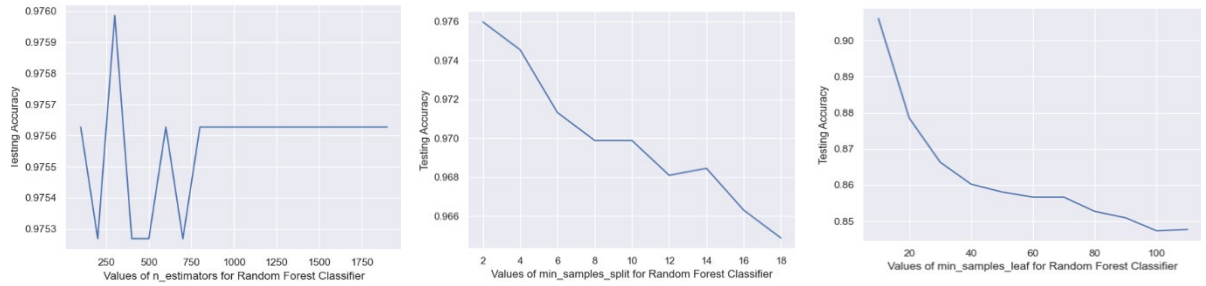
Families	Type	Samples #	Used Samples #	Percentage
Dowgin	Adware	3,410	3,410	100
DroidKungFu	Backdoor	549	549	100
Erop	Trojan-SMS	46	46	100
FakeAngry	Backdoor	10	10	100
FakeAV	Trojan	5	5	100
FakeDoc	Trojan	21	21	100
FakeInst	Trojan-SMS	2,172	1,997	91.9
FakePlayer	Trojan-SMS	21	5	23.8
Finspy	Trojan-SMS	9	9	100
Fjcon	Backdoor	20	20	100
Fobus	Trojan	4	4	100
Fusob	Ransom	1,277	1,270	99.4
GingerMaster	Backdoor	128	128	100
GoldDream	Backdoor	53	53	100
Koler	Ransom	69	4	5.8
Kuguo	Adware	1,199	10	0.8
Lotoor	HackerTool	333	15	4.5
Mecor	Trojan-Spy	1,820	1,820	100
MobileTX	Riskware	7	7	100
Penetho	Riskware	7	7	100
RuMMS	Trojan-SMS	402	350	87.1
SimpleLocker	Ransom	173	11	6.4
SlemBunk	Trojan-Banker	174	16	10.3
Youmi	Adware	1,301	76	5.8
Ztorg	Trojan-Dropper	20	3	15
Total/Average	-	14,437	10,960	76.9

As mentioned in above, the dataset consisted of 33 families and 10,960 applications that were not evenly distributed. With mean application number per family equal to 341 and median equal to 44, we can certainly conclude that this distribution is positively skewed and eventually unbalanced. Even though, we could use undersampling or oversampling to balance the dataset, it was decided to use the unbalanced data so that the results will be reproducible and reflect a real situation in which some variations of the malicious applications are more frequent than others for various reasons, e.g., the application code was publicly made available, etc. Oversampling the minority class can lead to overfitting and can also result in a disproportionate weight on the minority class, which may lead to biased predictions. Undersampling the majority class, on the other hand, can result in loss of valuable information and may lead to under-representation of important features in the model. Moreover, Random Forest is known for its ability to handle unbalanced datasets and does not require balancing of classes. Instead, it relies on the construction of decision trees on random subsets of the data, and by combining the predictions of multiple trees, it can effectively handle imbalanced data. Thus, since the dataset was unbalanced, parameters of the classification algorithm should be set accordingly. After thorough analysis for the best parameter combination with GridSearchCV method (Ranjan et al, 2019), we concluded that the best parameters for Random Forest Classifier model are those which are presented in Table 4 and are illustrated in bold characters.

Table 4: RF Initial Parameters

parameters	Option_1	Option_2	Option_3
bootstrap	**True**	False	-
oob_score	**True**	False	-
n_jobs	**-1**	None	-
warm_start	**True**	False	-
criterion	entropy	**gini**	-
max_features	auto	**sqrt**	-
min_samples_leaf	**1**	2	-
min_samples_split	**2**	10	-
n_estimators	50	**300**	1000

We can observe how accuracy fluctuates while n_estimators, min_samples_split, min_samples_leaf values' are changing in Figure 1 (a), (b) and (c) respectively.



(a) (b) (c)

Figure 1: Testing accuracy for values of n_estimators (a), min_samples_split (b) and min_samples_leaf (c)

We notice that the peak value of accuracy is detected for n_estimators = 300 and min_samples_split and min_samples_lead are negatively correlated with accuracy.

5. Results

The outcomes of performance metrics are shown in Table 5. The evaluation of the method is performed by considering detection parameters or in terms of accurately distinguishing among malicious families by considering accuracy, f1-score, recall and precision that are displayed in the Figure 2, and True Positive, True Negative, False Positive, False Negative ratio that are described in the confusion matrix generated by the predictions of RF model (Figure 2). Precision is defined as the probability of the proposed system correctly identifying the malicious application and is calculated based on True Label by predicted Labels. Recall refers to the number of total malicious apps of each family, which are labelled correctly to the family that they belong and can be determined by considering True Labelled by Total Labelled. Furthermore, accuracy of machine learning classifier has also been determined to demonstrate the effectiveness of Random Forest. In the experiment, the precision that was achieved was 98% with 97.52% recall and 97.56% f1-score. From the confusion matrix, we can deduce that 96.4% of applications were correctly classified (true labelled) to the family that they belong while only a 3.6% were misclassified (false labelled) according to the results.

Table 5: Evaluation metrics

Families	Precision	Recall	F1-score	Support
AndroRAT	1	1	1	12
Andup	1	0.909091	0.952381	11
Aples	1	1	1	5
BankBot	1	0.7625	0.865248	160
Bankun	1	0.882353	0.9375	17
Boqx	0.796875	0.944444	0.864407	54
Boxer	1	1	1	11
Cova	1	1	1	4
Dowgin	0.982415	0.982415	0.982415	853
DroidKungFu	1	0.992647	0.99631	136
Erop	1	1	1	11
FakeAV	1	1	1	1
FakeAngry	1	1	1	3
FakeDoc	1	1	1	5
FakeInst	1	1	1	541
FakePlayer	1	1	1	5
Finspy	1	1	1	2
Fjcon	1	1	1	5
Fobus	1	1	1	1
Fusob	1	1	1	317
GingerMaster	0.964286	0.870968	0.915254	31
GoldDream	1	1	1	12
Koler	1	1	1	1
Kuguo	1	0.666667	0.8	3
Lotoor	1	1	1	6
Mecor	1	1	1	454
MobileTX	1	1	1	2
Penetho	1	1	1	1
RuMMS	0.696	1	1	87
SimpleLocker	1	1	1	3
SlemBunk	1	1	1	4
Youmi	0.97	0.9	0.93	31
Ztorg	1	1	1	1
accuracy	-	-	0.975269	2,790
macro avg	0.980928	0.975254	0.970735	2,790
weighted avg	0.98005	0.975269	0.975666	2,790
Accuracy : 97.75%				

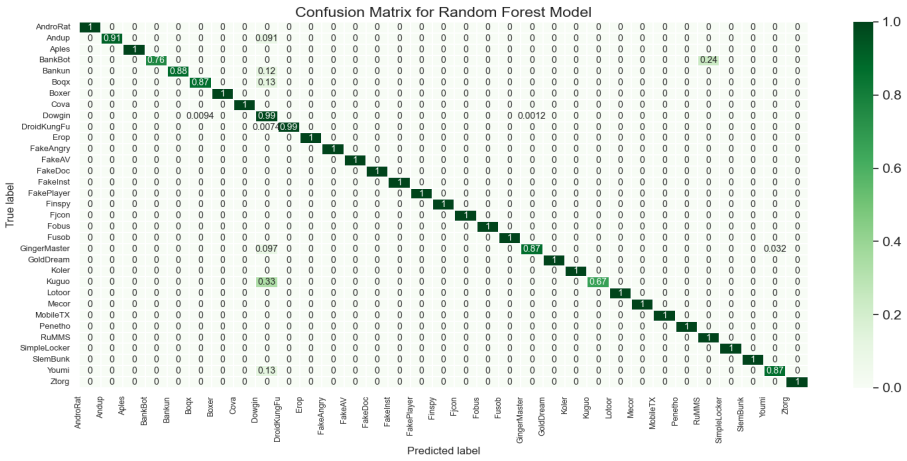
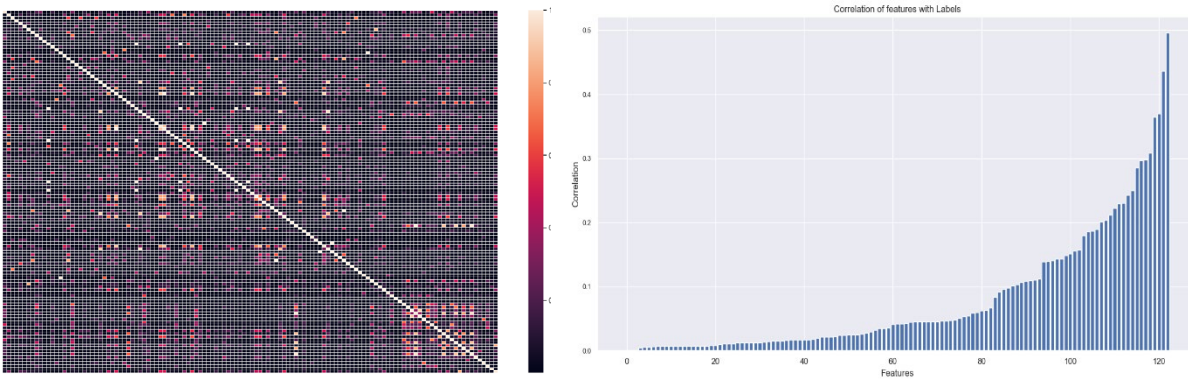


Figure 2: Confusion Matrix for RF Model

Thus, the proposed methodology proves to be highly efficient even though the dataset used is highly unbalanced. RF is a machine-learning method that generally works well with high-dimensional problems and allows for nonlinear relationships between features. In Figure 3 (a), where the heatmap depicts the intercorrelation of the labels and features, we can discern by the light colour that there are only a few instances in which features are highly correlated among them. Similarly in Figure 3 (b) we can see the correlation of each future with the applications' family labels.



(a) (b)

Figure 3: Intercorrelation heatmap (a) and correlation (b) of labels and features

Comparing our results with other published methodologies, we can see that our method outperforms several other methods in terms of higher precision, recall and f1-score. As for example, in Milosevic et al (2017), as it was mentioned, the researchers use permissions in a wide range of ensemble machine learning algorithms and the highest precision, recall and f1-score achieved was 89%. Pektaş and Acarman (2018) used a hybrid system for malware classification based on various features including permissions and they achieved 92% accuracy after 8 days of processing. Kang et al (2013) using the bytecode of the applications for classification with Decision Trees, they managed to score 94% of F-measure. Compared to the above-mentioned methods, our proposed model not only performs better and faster, but also uses a highly unbalanced dataset of 10,960 malicious applications, categorized in 33 families and requires only the permissions in order to classify these applications.

The proposed methodology even though it correctly classifies the malware applications in the appropriate family with very high precision, it has some limitations. If, for example, a new malware application of a known family will enter the test dataset then the method will work efficiently and the malware application will be classified to the correct family with high precision. However, a malware application which belongs to a new family or a family that was excluded from the training dataset will be classified to an existing family especially if the permissions are correlated with those of an existing one. Another limitation is that it is unable to classify a malware application if there are no permissions in the AndroidManifest.xml. We noticed in the given dataset, there are cases in which in the same family there was a small percentage of malware applications that they did not have any permission. In Table 1, it can be seen that FakeInst family has 5 varieties with 2,172 malware applications in

total, but from those only the 1,997 of them had AndroidManifest.xml file and could be used in the dataset. The remaining 197 malware applications could not be classified at all using the specific methodology. Finally, since Random Forests Classifier has been used to classify the malware applications which is a probabilistic classifier, statistical error is introduced. However, the error depends on the correlation between two trees in the forest which in our case is very low as depicted in Figure 2 and the strength of each tree in the forest (Breiman, 2001) which has been optimized based on the initial parameters selected.

6. Conclusions

Classification of malicious applications to malware families can be a difficult task due to their great diversity. There is an undeniable need for accurate and efficient categorization of malicious applications in order to determine the abilities of these applications and the possible ways that they may affect the victim. Therefore, we developed an effective mechanism that extracts information from the applications' permissions and by utilizing a well-established machine learning algorithm, Random Forest, it was possible to predict accurately in which family the sample application belongs to. The model was applied on a dataset of 10,960 applications distributed in 33 families and with hyperparameter tuning it reached an accuracy of 97.75%. Due to the simplicity and customizability of the method, it can be generalized to operate on other types of input data (receivers, intents, etc.) or even to tackle other types of classification problems such us to distinguish benign from malicious applications. In addition, the specific research work can be further expanded to other malicious data sources in order to further develop the classification model and achieve even better results. Furthermore, the technique can be enhanced by incorporating applications' features from dynamic analysis which could address the limitation of missing permissions in some malicious applications. These improvements we believe that can further enhance detection and classification of malicious Android applications and boost further research in this field.

References

Adebayo, O.S. and Abdul Aziz, N., (2015) Static Code Analysis of permission-based features for android malware classification using apriori algorithm with Particle Swarm Optimization.

Ahmad, I.N., Ridzuan, F., Saudi, M.M., Pitchay, S.A., Basir, N. and Nabila, N.F., (2017) Android mobile malware classification using a tokenization approach. In *The World Congress on Engineering and Computer Science* (pp. 271-285). Springer, Singapore.

Amer, E., (2021) May. Permission-based approach for android malware analysis through ensemble-based voting model. In *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)* (pp. 135-139). IEEE.

Mohamad Arif, J., Ab Razak, M.F., Awang, S., Tuan Mat, S.R., Ismail, N.S.N. and Firdaus, A., (2021) A static analysis approach for Android permission-based malware detection systems. *PloS one*, *16*(9), p.e0257968.

Biau, G., Devroye, L. and Lugosi, G., (2008) Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, *9*(9).

Breiman, L., (2001) Random forests. *Machine learning*, *45*(1), pp.5-32.

Hamming, R.W., (1950) Error detecting and error correcting codes. *The Bell system technical journal*, *29*(2), pp.147-160.

Huang, C.Y., Tsai, Y.T. and Hsu, C.H., (2013) Performance evaluation on permission-based detection for android malware. In *Advances in intelligent systems and applications-volume 2* (pp. 111-120). Springer, Berlin, Heidelberg.

Kaspersky (2022) "2021 Mobile threats report", [online], Kaspersky.com, https://www.kaspersky.com/about/ press-releases/2022_2021-mobile-threats-report-cybercriminals-forego-low-hanging-fruit-to-go-after-banking-and-gaming.

Li, J., Sun, L., Yan, Q., Li, Z., Srisa-An, W. and Ye, H., (2018) Significant permission identification for machine-learning-based android malware detection. *IEEE Transactions on Industrial Informatics*, *14*(7), pp.3216-3225.

Liu, Y., Wang, Y. and Zhang, J., (2012) New machine learning algorithm: Random forest. In *International Conference on Information Computing and Applications* (pp. 246-252). Springer, Berlin, Heidelberg.

Milosevic, N., Dehghantanha, A. and Choo, K.K.R., (2017) Machine learning aided Android malware classification. *Computers & Electrical Engineering*, *61*, pp.266-274.

Ranjan, G.S.K., Verma, A.K. and Radhika, S., (2019) K-nearest neighbors and grid search cv based real time fault monitoring system for industries. In *2019 IEEE 5th international conference for convergence in technology (I2CT)* (pp. 1-5). IEEE.

Panigrahi, R., Borah, S., Day, N., Babo, R. and Ashour, A.S., (2018) Classification and analysis of facebook metrics dataset using supervised classifiers. *Social Network Analytics: Computational Research Methods and Techniques*, *1*.

Pektaş, A. and Acarman, T., (2018) Ensemble machine learning approach for android malware classification using hybrid features. In *International conference on computer recognition systems* (pp. 191-200). Springer, Cham.

Petroc, T. (2023) "Smartphone mobile network subscriptions worldwide 2016-2028", [online], Statista.com, https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/.

Petroc, T. (2023) "Market share of mobile operating systems worldwide 2009-2022", [online], Statista.com, https://www.statista.com/statistics/272698/global-market-share-held-by-mobile-operating-systems-since-2009/

Ranjan, G.S.K., Verma, A.K. and Radhika, S., (2019) K-nearest neighbors and grid search cv based real time fault monitoring system for industries. In *2019 IEEE 5th international conference for convergence in technology (I2CT)* (pp. 1-5). IEEE.

Rovelli, P. and Vigfússon, Ý., (2014) PMDS: permission-based malware detection system. In *International conference on information systems security* (pp. 338-357). Springer, Cham.

Sachdeva, S., Jolivot, R. and Choensawat, W., (2018) Android malware classification based on mobile security framework. *IAENG International Journal of Computer Science*, *45*(4), pp.514-522.

Saudi, M.M., 2017. Mobile malware classification via system calls and permission for GPS exploitation. *International journal of advanced computer science and applications*, *8*(6).

Wei, F., Li, Y., Roy, S., Ou, X. and Zhou, W., (2017) Deep ground truth analysis of current android malware. In *International conference on detection of intrusions and malware, and vulnerability assessment* (pp. 252-276). Springer, Cham.

Cultural Influences on Information Security

Henry Collier[1], Charlotte Morton[2], Dalal Alharthi[3] and Jan Kleiner[4]
[1]Norwich University, Northfield, VT, USA
[2]University of Chester, Chester, UK
[3]The University of Arizona, Tucson, AZ, USA
[4]Masaryk University, Czech Republic

hcollier@norwich.edu
0818685@chester.ac.uk
dalharthi@arizona.edu
jkleiner@mail.muni.cz

Abstract: Humans are by far the weakest link in the information security chain. Many in the information security industry advocate for a technical solution to this problem. Unfortunately, technology does not hold the answer to solving the human problem. Instead, it is important to better understand the problem and find new ways of training individuals, so they have a better security mindset and make better security minded decisions. The security challenges associated with human factors have been widely studied in previous literature and different research groups. Prior research has shown that both human behavioural factors and social media usage factors can be used to better assess a person's susceptibility to cybercrime. We know that humans are multi-faceted beings who are swayed by many factors. In addition to behavioural factors and social media factors, humans are predisposed by cultural influences. This paper begins the process of understanding how culture influences a person's ability to make positive cybersecurity decisions in a world that is full of data being thrown at them. The end goal of this research is to use culture, along with behaviour and social media usage as new metrics in measuring a person's susceptibility to cybercrime. This information can then be used by information security practitioners and researchers to better prepare individuals to defend themselves from cyber threats. This paper is the start of the research process into how culture impacts a person's susceptibility to cybercrime. It shows the significance of identifying what specific aspects of culture impact how someone makes a decision. This can help mitigate social engineering attacks by better understanding the influencing factors which control an end user. The authors will continue their work on this project to develop new Information Awareness (IA) training programmes that work to modify an individual's behaviour, while taking into consideration their behaviours, social media usage and culture.

Keywords: cybersecurity, resilience, susceptibility, culture, social engineering, behaviours

1. Introduction

Humans are the front door to most computer networks and nefarious threat actors open this doorway with precision and ease. The nefarious actors open the human doorway easily because they understand that humans are the weakest link in the chain of information security (Collier, 2021) (Collier & Collier, 2020) (Hallas, 2018) (Hadnagy, 2018) (Schroeder, 2017). To open this door, the threat actors simply need to use the end user's behaviours against them (McIlwraith, 2006).

Historically speaking, researchers have tried to find a technical way to prevent social engineering. Add to this, industry has approached cybersecurity through technology and policy, neglecting the cultural and behavioural dimensions (Bernal, 2021). However, behaviour and awareness are some of the most critical issues that needs to be addressed in best practice guidelines and protective tools (Zwilling, et al., 2020). Unfortunately, people circumvent every technical measure created to protect them because these methods make their lives more difficult (Collier, 2021) (Collier & Collier, 2020) (McIlwraith, 2006) (Zinatullin, 2016). A 2019 study investigated the reasons people most often resist cybersecurity policies; it found that if the policies are overly complex or punitive in nature, the end user will work to find a way around the policy, while giving the appearance of compliance (Mehri & Ahluwalia, 2019). Moreover, when people are subjected to rules and norms that are not robustly justified to them, their reflexivity towards these rules suffers, or they tend not to follow them at all (D´Agostino et al., 2021). Coles-Kemp et al. (2018) predicted a similar ramification if people´s desires and needs are not taken into consideration. These are some of the underlying aspects of the social contract theory which

has been increasingly applied to cybersecurity (see, e.g., Abrams et al., 2016; Liaropoulos, 2020) and which is closely tied to culture.

Research into the reasons why a person makes a poor information security decision has increased during the last five years, but more needs to be done. One of the perspectives that has been considered is working on making the security mindset more prevalent amongst end users. A cybersecurity mindset was defined by (Dutton, 2017) as, "a set of attitudes, beliefs and values that motivate individuals to continually act in ways to secure themselves and their network of users, such as by acquiring technical skills, new practices or changing their behaviour online". By choosing a framework approach to cybersecurity that is bottom up, cybersecurity practices will be socially supported (Dutton, 2017), rather than a top-down approach, which is often associated with a fear driven concept. Fear is negative in nature and does not result in long term positive outcomes. Fear feeds into the concept of a toxic workplace and toxicity often results in employees second guessing themselves (Hassandoust & Techatassanasoontorn, 2018) (Mehri & Ahluwalia, 2019) (Collier, 2021). Research has shown that using punishment as a means of behavioural modification is ineffective compared to positive reinforcement (Dynes, et al., 2020) (Durrant & Ensom, 2012). To develop a cybersecurity mindset, it is important not to create a culture of fear (Dutton, 2017) but for organisations to think of the human layer of defence, differently. Instead of putting end users in a position where they might try to conceal a potential breach out of fear of reprisal, the cybersecurity community needs to shift its perspective and work to identify breaches sooner by encouraging employees to notify the information security team of a suspected breach, or if they have fallen for a social engineering scheme. If a security breach is identified sooner, the security team can implement a recovery strategy quicker, stopping the damage faster. Stopping the exfiltration of data quicker will result in less information being lost, and reduces the overall damage done by the threat actor. This lower level of data loss will not only reduce the financial burden of the company that was targeted, but also reduces the risk of long-term negative impacts on the customers that the company serves.

Zimmerman & Renaud (2019) proposed a mindset called 'Cybersecurity, Differently'. This approach changes the human from being the weakest link in cybersecurity to being part of the solution, looking through the lens of the socio-technical system (STS). STS considers the interconnectedness of hardware, software, people, and communities. The human is widely blamed by governments and industry as the weakest link in cybersecurity. For example, IT would be blamed for a successful cyber-attack if a hacker exploits an unpatched system, while the user would be blamed for not paying enough attention during the training sessions if they fell for a phishing email (Zimmerman & Renaud, 2019). In both cases, the blame is on the human, not the technology.

The current defence against the human problem in cybersecurity is the annual cybersecurity awareness training and assessment used by most companies to "train" their employees on cyber issues. Cybersecurity awareness training programmes teach end users about the risks associated using networked systems. It is imperative that organizations ensure their employees understand the risks of cybercrime, and to be prepared to not fall victim to it (Alharthi & Regan, 2021). Unfortunately, knowledge is not enough because the current methods of IA training do not change a person's behaviour, nor do they require a person to learn the material, the individuals are only required to achieve a certain score on the quiz (Schroeder, 2017) (Collier, 2021). Acknowledging the concepts that end users need to be aware of is not enough to make a significant difference (Schroeder, 2017). People also face the problem of information overload and fatigue (Hallas, 2018). Cybersecurity awareness can be defined as 'the degree of understanding of users about the importance of information security and their responsibility and acts to exercise sufficient levels of information security control to protect the organisation's data and networks' (Shaw, et al., 2009). Hackers will seek out the most vulnerable end user (Zwilling, et al., 2020). 'Netizens,', people who use the internet, lack an acceptable level of awareness of the different types of cybersecurity threats, in this study 'threats' is defined as 'cyber hazards' (Zwilling, et al., 2020). These go from the very simple spam emails or texts and calls to organised cyber-crime 'people'. One of the recommendations is that training programmes should be developed from an international position related to an individual's behaviour, rather than based on local or cultural expressions (Zwilling, et al., 2020).

The result of a breach associated with a person falling victim to cybercrime is the termination of the person who fell victim from their employment (Collier, 2021). When an organization is breached, regardless of the reason why, one of the first victims is the Chief Information Security Officer (CISO) and sometimes the entire C-suite is impacted. Target, Capitol One, Equifax, Uber, and JP Morgan each suffered breaches in which the CISO, CIO and/or CEO were terminated due to mishandling of the incident or simply because the breach occurred (Swinhoe, 2020). If one considers that it is not if, but when, an organization falls victim to a breach, then the idea

that an organization immediately terminates an employee or member of the executive team is counterproductive in the long term. More needs to be done to understand why someone becomes a victim to cybercrime and to identify and develop new ways of defending against such breaches. Collier, in his 2021 study used additional metrics to evaluate a person's susceptibility to social engineering (Collier, 2021). The results showed that when factors like behaviours and social media usage were calculated into a susceptibility algorithm, it was possible to assess someone's susceptibility to cyber-crime in a more finite manner (Collier, 2021). However, behaviours and social media usage factors are only a couple of the metrics that need to be assessed when calculating a person's susceptibility level. People are multifaceted beings, influenced by many factors, most of which are unmeasurable. The goal of research needs to be to identify which influential factors impact the decision-making process and affect security minded decisions. If it is possible to better understand the people who are falling victim to cybercrime and what influencing factors impacted their cognitive ability and decision-making process, then it will be possible to develop new methods and techniques to prevent future breaches.

This paper is a first look into how culture plays a role in a person's decision-making process and whether culture impacts the security mindset of individuals. As the culture´s role in an individual´s decisions regarding cybersecurity has not yet been investigated much by scholarly literature, the aim here is to employ an exploratory investigation (as conceptualised by Stebbins, 2001) to lay down the foundations for further empirical and theoretical works. The remainder of this paper is structured as follows. Section 2 highlights how culture plays a role in the decision-making process and how it impacts the security mindset. Cultural influence on cybersecurity decision-making is addressed in Section 3. The paper then addresses a proposed solution to the research problem in Section 4 and concludes by addressing future research venues in Section 5.

2. Culture/decision-making process/security mindset

To understand how culture plays a role in the decision-making process and how it impacts the security mindset, we need to understand what culture is, what the decision-making process is, and what a security mindset is. Once these aspects are better understood, it is important to identify how they are linked and how one influences another.

2.1 Culture

Simonson, et al., (2000) propose that "cultures endow individuals with different rules or principles that provide guidance for making decisions, and a need to provide reasons activates such cultural knowledge" and imply that culture is something that is manipulated by societal rules, which means that culture is learned and changes over time. Since culture is learned, it is consequently guided by societal rules (Whiten, et al., 2011). Culture is by default abstract or fuzzy in nature (Causadias, 2020). Due to the ambiguous nature of culture, it is difficult to define exactly what culture is (Causadias, 2020). Culture is sophisticated and influenced by a set of distinct and similar individuals who are connected on various levels, all related to behavioural expectations, religious beliefs, morals, values, language, symbolism, and ethics. In other words, culture is social in nature (Causadias, 2020). One important thing that needs to be understood is that "culture is not an insignificant or marginal object" (Breznik, 2013).

Müller, et al. (2009) notes that "numerous studies have been conducted on cultural differences and commonalities, showing the differences in values and behaviour of people from different national cultures." Culture can be placed in one of two modalities—individualism and collectivism (Müller, et al., 2009) (Darwish & Huber, 2003) (LeFebvre & Franke, 2013). Individualism is where people are more concerned with themselves and their family members, and less concerned with others within their community (Darwish & Huber, 2003). Conversely, collectivism is where people are more concerned about their community, versus themselves or close family members (Darwish & Huber, 2003). This paper does not judge either individualism or collectivism as being the right way to approach culture, but rather reinforces that culture is not simple in nature. From an information security perspective, individualism and collectivism impact the decision-making process differently and could result in a vastly different outcome. When looking at culture from a worldly perspective, North America and Europe tend to be individualistic cultures, while East Asia and Middle Eastern countries tend to be collective cultures (Darwish & Huber, 2003).

The current statistical information regarding which countries have the most incidents of cybercrime vary depending on the source and is therefore unreliable when assessing if individualism or collectivism is better when it comes to the information security mindset. What is consistently seen when looking at this information is that these incidents span across both individualistic and collective cultures, so neither culture is immune. With most cyber incidents involving social engineering, it is imperative that we understand how culture plays a role in individuals becoming victims to social engineering (Collier, 2020) (Bada, et al., 2019) (Dupuis & Khadeer, 2016).

What we have learned so far about culture is that it is something that is complex in nature, learned through experiences and social group interactions, it is flexible, and most importantly, culture is not insignificant. This creates a problem when trying to determine a method of assessing the risk associated with culture and the information security decision making process.

2.2 Decision-making process

The decision-making process is a process whereby an individual makes a choice based on available data and assessing multiple options. The decision-making process is influenced by knowledge, information, and resources, to include past experiences and skills possessed by the decision maker. The decision maker's psychology, behaviours, and emotional state, also affect the decision maker (Collier, 2021). The basis of the decision-making process is that there is a goal, there are choices and there are selection criteria being considered (Wang & Ruhe, 2007).

Decision-making is part of the cognitive process that separates humans from other mammals. Humans can take detailed amounts of data into consideration and have the ability to think beyond the moment. The decision-making process occurs in both the conscious mind and subconscious mind (Wang & Ruhe, 2007). When conducted in the conscious mind, the individual making the decision will apply a logical approach to the decision-making process. However, in the subconscious mind, behaviours and emotions get involved and the decision-making process becomes muddled and loses some, if not all, of its logic, which may result in a bad decision. For example, logic would dictate that an adult human being could easily step on or remove an insect from their house without harm. However, if the human suffers from entomophobia, logic goes out the window and the decision-making process is heavily influenced by the fear of the insect. Fear, like many other human emotions and behaviours, can be significant influencing factors when it comes to the decision-making process.

Decision-making, like culture, has been studied for centuries by philosophers, but has a shorter history with experimental psychology (Slovic, et al., 1988). Although many different models have been developed over the decades of research that attempt to define the decision-making process, these models are not predictive in nature. They might explain how a decision was made, but when applied to a new decision, there is no data that supports the models that can predict the outcome of the decision with statistical significance (Slovic, et al., 1988). This is because each decision is unique and shaped by many things (Slovic, et al., 1988). For example, when two people are given the same data, do they come up with the same decision every time? Of course not, because each of the decision-makers is a unique individual with different psychological make ups, emotions, knowledge, desires, and cultural influences. It is these unique attributes which make humans wonderful, but at the same time puts them in a position of risk when it comes to the information security decision making process. When you take into consideration all the inputs that impact a decision, it is clear why and how social engineers are as successful as they are. Humans are flawed, as is their security-minded decision-making process.

2.3 Security mindset

The security mindset is a state of mind whereby an individual who is deciding, consistently considers the security outcome of that decision. Since the advent of the computer and the Internet, humans have become more connected than ever. This constant connectedness comes with many benefits, but also many detriments. From a business perspective, the lightning speed of the Internet allows for the near immediate transferal of data, this input of data in such an available manner makes business decisions faster and more accurate. However, this comes with a flaw and that is there is a ton of information available at the touch of a few keys. Much of this data is meant to be kept secret, but often it falls into the hands of nefarious actors. Since humans are the front line of any organization's network, one would think they are well prepared to defend the network and to secure the data. This concept is about as far from the truth as one can get. Humans are imperfect, and their weaknesses are the reason why humans are the number one entry point into a secure network.

If an organization works to foster a security mindset within the population of employees, then it is working to build a more secure border to its network. However, it is important to realize that if not properly developed, the security mindset can be utilized too much, or not utilized enough (Dutton, 2014). The cybersecurity industry understands that more needs to be done to focus people's attention on "attitudes, beliefs and practices" (Dutton, 2017). Unfortunately, the cybersecurity industry and most cybersecurity researcher's response is to push for the development of new technical measures to close the hole in the human firewall (Collier, 2021). As each new technical measure is created, a human will figure a way around it, to circumvent it because it is inconvenient. Instead of finding a technical measure to solve the human problem, we propose understanding the human better to strengthen the human firewall. One of the most important things that we can do to solve this issue is to work to better understand why someone becomes a victim of cybercrime. From here, we need to work to have end users on our systems take ownership of the security and work to develop a security- mindset.

"The security mindset is not only essential but critical for all individuals who design/develop/deploy/upkeep/use digital systems" (Siraj, et al., 2021). Instead of hoping that your company's cyber team is going to catch all attacks and exploits, it is time that end users take responsibility for their own actions. Developing a security mindset has the potential to go beyond the existing technical measures and redefine how a person behaves (Hann, 2022).

3. Cultural influence

From a security perspective, there are three main types of threats to a network. The first is the malicious outsider threat (MOT), that black hat hacker/nefarious threat actor who is trying to get into your network and steal your data or encrypt it. The second is the malicious insider threat (MIT), that disgruntled employee who wants to damage the company because their egos have been bruised and they feel as though they have been wronged by the people they work with or the organization itself. These first two are not what this research is about. Rather, this research is about the third type, the non-malicious insider threat (NMIT) who doesn't want to hurt their organization but does because they fall victim to social engineering. The NMIT threat is the threat that can be prevented by understanding the human element better and then helping them develop a better security mindset.

Collier (2021) conducted research into the NMIT and worked to better understand how human behaviours and social media usage impacted the NMIT. As part of his research, Collier determined that there are 128 different human behaviours which lead to a person making a poor information security decision, resulting in them becoming a victim of social engineering. Collier (2021) developed a new Dynamic Adaptable Information IA Training Assessment tool that utilized questions based on these behaviours and questions related to social media usage to assess one's susceptibility to cybercrime in a more precise manner. This study is an extension of what Collier did with his Dynamic Adaptable IA Training Assessment Tool in that we are working to determine how culture influences a person's information security decision-making process and to see if we can assess and measure it effectively.

In section 2.1, we demonstrate that culture is something that is integrated into a person from birth. Like morals and ethics, culture influences the decision-making process. For example, if a person's culture is that of collectivism, then the person will tend to make decisions that benefit their community over themselves. While on the contrary, someone from an individualistic culture, will tend to make decisions that benefit themselves, over their community at large. Regardless of which form of culture one is associated with, it is clear culture impacts the decision-making process. The question then becomes how does culture influence the information security decision-making process, and can it be measured? We also need to ask if it is possible to develop a set of questions that can be used to generalize culture or is culture so unique that it needs to be applied in a homogeneous manner based on where the tool is being used.

There is no global IA training assessment tool that truly assesses the unique susceptibility each person has. Currently things are artificially grouped together to meet a requirement to train and assess. The current methodologies do not consider human behaviours, culture, or external influencing factors like social media. The cliché about insanity from Albert Einstein states doing the same thing over and over and expecting a different result is insanity may be one of the most overused quotes in history. However, when it comes to the way the cybersecurity industry approaches IA training and assessment, the statement isn't far off. The current IA training and assessment models do nothing to truly change a person's behaviour, and part of the reason is because these methods were not designed around the person, their behaviours, or cultural influences.

To be more effective at changing a person's behaviour, IA training and assessment needs to be reformatted to include psychological techniques like Cognitive Behavioural Therapy (Collier, 2021). Furthermore, these training and assessment programmes need to utilize behavioural and cultural data sets to make the training more international in nature (Zwilling, et al., 2020) (Collier, 2021).

4. The social contract theory, governance, polity, and culture

The social contract theory (SCT) and cybersecurity governance are only a step away from decision-making. In fact, these two often overlap. Note that governance does not need to be solely connected to the state level but also to the organizational one. The SCT represents centuries of political philosophy thinking, from ancient first fruits through Hobbes, Rousseau, or Locke to Rawls, Buchanan or Scanlon. Despite its many branches and particular theories, its basic model comes down to representatives (e.g. politicians or other decision-makers) choosing rules, norms, or principles in a deliberative setting for individuals who are to be guided by these rules and adhere to them. Both these groups must share reasons and norms for the sake of compliance and reflexivity (D´Agostino et al., 2021). Moreover, such principles must be robustly justified and reach a sufficient level of publicity (Gaus, 2021).

The justification process can vary significantly based on many factors. One of them is polity which Carr (2007) links with political (or institutional) culture. Within, the political system plays a huge role. To demonstrate, let us stay with democracy, which can be broadly grasped as a process (procedure) or content (generated rules and laws). If democracy as a procedure is sufficiently justified, its products (contents) are, too (Swift, 2014). In such a context, we use the term legitimacy. The concrete democratic justification then stands on rather different foundations than the undemocratic one, for instance, the Chinese communist or Iranian theocratic polities. Moreover, the contractualist strain of the SCT represented by Kant, Rawls, Scanlon and others cannot be used outside the democratic scope as they emphasize freedom, justice, property rights etc. (Lessnoff, 1990).

Finally, the concept of the Four Internets developed by O´Hara and Hall (2021) loosely connects the recognition of democratic and non-democratic regimes as a part of the culture and the cyber realm. The authors distinguish four internet governance approaches - Silicon Valley´s open, [Washington] DC´s commercial, Brussels´ burgeois, and Beijing´s authoritarian, plus Moscow´s spoiler model. Hence, even within the democratic scope, we have sub-clusters that emphasize openness, commercial exploitation, data and user protection. The similar applies to undemocratic ones.

In sum, governance on state and organizational levels is closely tied with or even constitutive of culture and probably vice versa. To theoretically grasp these phenomena, we employ the social contract theory - one of the most potent branches of political philosophy that has been increasingly applied to the cyber(security) realm. Here, we must reckon with political systems as well as they provide different foundations for one of the core SCT concepts - justification of rules, norms, and principles. This applies not only to the basic distinction between democratic and undemocratic ones but also to a deeper recognition of various strains located within these two traditions. Finally, by adding this discussion to our mix, we are then able to formulate cybersecurity governance implications.

5. Moving forward

We must understand how culture plays a role in the information security decision-making process as it relates to changing how the end user works to protect their organization's data. To do this, we need to identify what specific aspects of culture impact how someone makes a decision. The first step was to understand what culture is, and this part of the study has given us valuable information about culture and how it is going to be unique depending on where in the world someone is from. These differences can be between countries, between regions, and could even be different between villages/towns. The following steps should thus test these confounding variables and look for other ones via both qualitative and quantitative empirical investigations. The complexity that culture is, is certainly going to make determining how culture impacts the information security mindset a bit more difficult. However, this research needs to be completed in order for the fight against the social engineers to be successful.

Now that we have a better understanding of what culture is we will work to establish a series of questions that demonstrate how culture influences the decision-making process. From these questions, we will work to refine

the questions to specific questions that impact the information security decision making process. Once the questions have been developed, we will test their efficacy using the Dynamic Adaptable IA Training Assessment Tool developed by Collier (2021) to determine if culture can be used to predict someone's susceptibility to cybercrime more accurately.

The current susceptibility algorithm used by Collier in his Dynamic Adaptable IA Training and Assessment Tool will be modified to include culture as one of its quantitative values. The proposed algorithm will be (QrTotal/Hb+SM+C)*1000)/2 where QrTotal is the average of the responses to the seven topics of cybersecurity questions, where Hb is the score associated with the human behavioural questions, where the SM score is related to the social media usage questions and where C is the value assigned based on the responses to the cultural assessment questions. The results of this algorithm will fall within the scale of 0-100, with 0 having the highest risk of becoming a victim, and 100 having the least risk.

Upon completion of this test, the results will then be used as part of an interdisciplinary team effort to develop new IA training programmes that work to modify an individual's behaviour, while taking into consideration their behaviours, social media usage and culture.

Acknowledgements

We would like to thank Norwich University student, Emily Collado for her work as a research assistant on this project.

References

Abrams, B., Barrack, S., Bew, R., Brown, J., Buonomo, J., Clinton, L., Connelly, J., Cotton, A., Crisp, D., Daly, K., Doucet, T., Estlick, D., Flannery, B., Fleming, M., Hermanson, S., Humbert, L., Khona, M., Jainchill, C., McAlum,

Alharthi, D. & Regan, A., 2021. *Social Engineering InfoSec Policies (SE-IPs).* Zurich, CICT.

Bada, M., Sasse, A. M. & Nurse, J. R., 2019. Cyber Security Awareness Campaigns: Why do they fail to change behaviour?. *CoRR,* Volume abs/1901.02672.

Breznik, m., 2013. And Culture?. *Journal of Arts and Humanities / Journal of Arts and Humanities,* 7(1), pp. 21-29.

Causadias, J. M., 2020. What is culture? Systems of people, places, and practices. *Applied Developmental Science,* 24(4), pp. 310-322.

Coles-Kemp, L., Ashenden, D., & O'Hara, K. 2018. Why should I? Cybersecurity, the security of the state and the insecurity of the citizen. Politics and Governance, 6(2), 41–48. https://doi.org/10.17645/pag.v6i2.1333

Collier, H., 2020. *Social Media: A Social Engineer's Goldmine.* Reading, Academic Conferences and Publishing International Limited.

Collier, H., 2021. *Enhancing Information Security by Identifying and Embracing Executive Functioning and the Human Behaviours 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 Behaviours Related to Susceptibility.* Copenhagen, AIRCC Publishing Corporation, p. 5.

D´Agostino, F., Gaus, G., & Thrasher, J. 2021. Contemporary Approaches to the Social Contract. https://plato.stanford.edu/archives/win2021/entries/contractarianism-contemporary/

Darwish, A.-F. E. & Huber, G. L., 2003. Individualism vs Collectivism in Different Cultures: a cross-cultural study. *Intercultural Education,* 14(1).

Dupuis, M. & Khadeer, S., 2016. *Curiosity Killed the Organization: A Psychological Comparison between Malicious and Non-Malicious Insiders and the Insider Threat.* Boston, s.n.

Durrant, J. & Ensom, R., 2012. Physical Punishment of Children: Lessons from 20 Years of Research. *CMAJ,* 184(12), pp. 1373-1377.

Dutton, W. H., 2014. *Fostering a Cybersecurity mindset.* [Online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2490010 [Accessed 30 01 2023].

Dutton, W. H., 2017. Fostering a cyber security mindset. *Journal of Internet Regulation,* 6(1).

Dutton, W. H., 2017. Fostering a cyber security mindset. *Internet Policy Review Journal on Internet Regulation,* 6(1).

Dynes, M. et al., 2020. Impact of Education about Physical Punishment of Children on the Attitudes of Future Physicians,. *Children's Health Care,* 49(2), pp. 218-231.

Hadnagy, C., 2018. *Social Engineering: The Science of Human Hacking.* Indianapolis: Wiley.

Hallas, B., 2018. *Rethinking the Human Factor.* s.l.:The Hallas Institute.

Hann, H., 2022. Using Complexity Theory to Identify K-12+ Pedagogical Misalignment with a Security mindset. *Journal of The Colloquium for Information Systems Security Education ,* 9(1).

Hassandoust, F. & Techatassanasoontorn, A. A., 2018. *Understanding User's Information Security Awareness and Intentions: A Full Nomology of Protection Motivation Theory.* Yokohama , s.n.

LeFebvre, R. & Franke, V., 2013. Culture Matters: Individualism vs. Collectivism in Conflict Decision-Making. *Societies,* Volume 3, pp. 128-146.

Liaropoulos, A. 2020. A Social Contract for Cyberspace. Journal of Information Warfare, 19, 1–11.

McIlwraith, A., 2006. *Information Security and Employee Behaviour: How to Reduce Risk Through Employee Education, Training and Awareness.* Burlington : Gower Publishing Company.

Mehri, M. I. & Ahluwalia, P., 2019. Examining the Impact of Deterrence Factors and Norms on Resistance to Information Security Systems. *Computers in Human Behaviour,* Volume 92, pp. 37-46.

Müller, R., Spang, K. & Ozcan, S., 2009. Cultural differences in decision making in project teams. *International Journal of Managing Projects in Business,* 2(1), pp. 70-93.

O'Hara, K., & Hall, W. 2021. Four Internets. Oxford University Press. https://doi.org/10.1093/oso/9780197523681.001.0001

Schroeder, J., 2017. *Advanced Persistent Training: Take your Security Awareness Programmes to the Next Level.* Edinburgh: Apress.

Shaw, R. S., Chen, C. C., jarros, A. L. & Huang, H.-J., 2009. The impact of information richness on information security awareness training effectiveness. *Computers & Education,* 52(1), pp. 92-100.

Simonson, I., Morris, M. & Briley, D., 2000. Reasons as Carriers of Culture: Dynamic versus Dispositional Models of Cultural. *Journal of Consumer Research,* 27(2), pp. 157-178.

Siraj, A. et al., 2021. Is there a Security mindset an Can it be Taught?. *CODASPY '21: Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy,* pp. 335-336.

Slovic, P., Lichtenstein , S. & Fischhoff, B., 1988. Decision Making. In: Wiley, ed. *Steven's Handbook of Experimental Psychology.* New York: Wiley.

Stebbins, R. A., 2001. *Exploratory research in the social sciences.* SAGE.

Swinhoe, D., 2020. *CSO.* [Online] Available at: https://www.csoonline.com/article/3510640/7-security-incidents-that-cost-cisos-their-jobs.html [Accessed 14 01 2023].

Wang, Y. & Ruhe, G., 2007. The Cognitive Process of Decision Making. *International Journal of Cognitive Informatics and Natural Intelligence.*

Whiten, A., Hinde, R. A., Laland, K. N. & Stringer , C. B., 20011. Culture evolves. *Philosophical Transactions of the Royal Society B,* 366(1567).

Zimmerman, V. & Renaud, K., 2019. Moving from a 'human-as-problem" to a 'human-as-solution" cybersecurity mindset. *International Journal of Human-Computer Studies,* Volume 131, pp. 169-187.

Zinatullin, L., 2016. *The Psychology of Information Security.* Cambridgeshire: IT Governance Publishing.

Zwilling, M. et al., 2020. Cyber Security Awareness, Knowledge and Behaviour: A Comparative Study. *Journal of Computer Information Systems,* 62(1), pp. 82-97.

Designing a high-fidelity Testbed for 5G-based Industrial IoT

Diogo Cruz, Tiago Cruz, Vasco Pereira and Paulo Simões
University of Coimbra, CISUC, DEI, Portugal

diogocruz@student.dei.uc.pt
tjcruz@dei.uc.pt
vasco@dei.uc.pt
psimoes@dei.uc.pt

Abstract: With the rise of the Industrial IoT (Internet of Things) and Industry 4.0 paradigms, many control and sensor systems used for IACS (Industrial Automation and Control Systems) have become more complex, due to the increasing number of interconnected field devices, sensors and actuators often being geographically spread across large areas. Supporting these increasingly sophisticated networked scenarios calls for the involvement of telecommunications and utility providers to better support Machine-to-Machine (M2M) communications and infrastructure orchestration, for which 5G technology is considered a perfect match.

Nowadays, such 5G networks empower solutions both for consumer and for industrial IoT scenarios, providing the capacity and the means to seamlessly connect a massive number of gadgets and sensors, with diverse data rate requirements, low latency, and low power consumption. Part of this flexibility is also due to the nature of the 5G Service Architecture (SA), which is based on a microservice concept, dividing its core through multiple functions, allowing it to horizontally scale in a flexible way. Furthermore, the 3GPP specifications encompass specific support for verticals by means of slicing and 5G LANs, paving the way for a paradigm shift in terms of the relationship between service, telecom, and operational infrastructure tenants. However, such benefits come at the cost of extra complexity and, consequently, an increased vulnerability surface. This calls for further research focused on improving 5G infrastructure management, service integration and security, which cannot be safely undertaken in production environments, thus motivating the development of suitable 5G testbeds.

This research work, which was developed in the scope of the POWER and Smart5Grid P2020 projects, addresses the creation of a high-fidelity environment for 5G-related research, which encompasses a gNodeB and 5G core, together with emulated User Elements (terminal devices) and IoT nodes (in this specific case, Programmable Logic Controllers), constituting a 5G Industrial IoT scenario designed for development and validation of new solutions, security research, or even advanced training purposes. The entire infrastructure is supported via container orchestration technology, providing enhanced scalability and resilience characteristics.

Keywords: 5G, Industrial IoT, Testbeds, Cyber-ranges

1. Introduction

The fifth-generation standard for broadband cellular networks, often designated as 5G, was devised to provide faster internet speeds, lower latency, and more capacity than previous generations. It has the potential to be used in various applications, such as self-driving cars, remote surgery, virtual reality and more. 5G is also considered an enabler for evolved Internet of Things (IoT) use cases, allowing to create and support networks of intelligent devices and appliances, that are embedded with sensors, software, and other technologies that enable them to collect and exchange data with each other and with other systems. The IoT paradigm has the potential to revolutionize a wide range of industries, including healthcare, transportation, and manufacturing, by enabling the development of new products and services that can improve efficiency, productivity, and safety. Unsurprisingly, the integration of 5G technology with IoT is widely regarded as an opportunity to create new and innovative products and services, taking advantage of the faster, more responsive, and reliable connections provided by 5G networks, which can also scale to support a large number of devices and data-intensive applications.

This paper describes the design of a testbed for research of 5G-driven Industrial IoT, developed within the scope of the POWER (POWER, 2020) and Smart5Grid (Smart5Grid, 2021) research projects. This testbed includes a 5G Core (5GC), Programmable Logic Controller (PLC) instances, User Equipment (UE) instances, and Next Generation Radio Access Network (NG-RAN) emulation components, supported by means of a container orchestration framework infrastructure. This testbed is used to implement a 5G-enabled Supervisory Control And Data Acquisition (SCADA) Industrial IoT scenario using several PLCs, designed with two objectives in mind: to add a practical element to the research, and to help validate the testbed itself. Overall, this testbed aims to explore the potential of these technologies and their ability to drive progress in various fields.

The rest of this paper is organized as follows. Section 2 presents a quick primer on 5G technology. Section 3 introduces the key technologies that enable/support the testbed. Section 4 delves into the testbed from a user

perspective, examining its structure and network topology. Section 5 describes how the testbed is used to implement a simple but representative Use Case, and Section 6 concludes the paper.

2. A quick primer on 5G technology

In the following subsections, we will explore the system overview of 5G, including its architecture based on the Service-Based Architecture (SBA) framework. We will also delve into the topic of 5G verticals and 5G Local Area Networks with support for the Industrial Internet of Things (IIoT).

2.1 System Overview

The current 5G specification is the culmination of the work started by the International Telecommunication Union in 2015, with the publication of the International Mobile Telecommunications – 2020 (IMT-2020) standard (ITU, 2015), which listed the requirements for future 5G networks. Following the availability of these requirements, the 3rd Generation Partnership Project (3GPP) (3GPP, 2022) started developing the technical specifications to implement the IMT-2020 standards in mobile communication networks. In 2018, 3GPP released its first package of technical specifications for standalone 5G networks, known as Release 15 (3GPP TSG, 2019). This release included the full batch of 3GPP specifications for 5G, introducing a New Radio (NR) access technology and other important concepts, geared towards improved reliability, increased modularity, and faster response times.

Release 16 (3GPP TSG, 2020), which was functionally frozen in December 2020, specified the second phase of 5G deployment. Release 16 built upon the foundations established in Release 15 to introduce new features and capabilities to further improve the performance of 5G networks. Some of the key features of Release 16 included:

- Enhanced Mobile Broadband: aimed at improving the quality of mobile broadband services by increasing the data rates and capacity of 5G networks.
- Ultra-Reliable Low Latency Communications: providing low latency and high reliability for applications that require real-time communication, such as remote surgery or autonomous driving.
  Massive Machine-Type Communications: allowing to support a large number of devices with low data rate requirements, such as sensors and other IoT devices.

Release 16 presented enhancements to the 5G Core network designed to increase its flexibility and ease of maintenance and provisioning, including support for network slicing, which allows for the simultaneous deployment and use of multiple Core Networks, each one specialized in providing a specific set of services and/or subscribers. It also introduced Network Function Virtualization (NFV) and support to edge computing, the latter targeted at bringing computational power closer to the end-user to reduce the response time of the network for applications that require low latency, such as virtual reality, autonomous driving, or evolved IoT.

Release 17 (3GPP TSG, 2022), functionally frozen in March 2022, brings a range of enhancements that focus on increasing capacity and coverage, improving latency, reducing power consumption, and expanding support for a wide range of devices and applications. Some of the key areas of improvement include further development of massive MIMO, coverage enhancements for diverse deployments, power savings for mobile devices, spectrum expansion to higher frequency bands, and enhanced support for ultra-reliable low-latency communication and private networks. Additionally, it includes new capabilities such as support for reduced capability devices (important in the scope of IoT), support for non-terrestrial networks, expansion of sidelink capabilities, and improved precise positioning. These enhancements push the technology boundaries on many fronts, making 5G more versatile and efficient.

Next, we will examine the architecture of 5G and explain the various components that make up this complex system. We will also discuss the functions and capabilities of these components, and how they contribute to the overall performance of 5G networks.

2.2 Architecture

As previously referred, the 5G architecture is an SBA. An SBA is a design approach for building systems in which the architecture elements are defined in terms of Network Functions (NFs) rather than by traditional network entities. In an SBA, any given NF offers its services to all other authorized NFs and/or to any “consumers” that are permitted to use these provided services. This approach offers modularity and reusability, as it allows for

the creation of flexible and scalable systems that can be easily modified or expanded as needed to support new features and services. (3GPP, 2022).

In the context of 5G, the 3GPP has adopted an SBA framework for the 5GC. The 5GC is the part of the 5G network that controls the communication between the User Equipment (UE), itself composed of a Mobile Station and a Universal Subscriber Identity Module (USIM), and the data network. It is formed of several NFs that work together to provide various services to the UE. These NFs can be grouped into two main categories: the User Plane, which handles user data transport, and the Control Plane, which handles network signalling and control.

In 5G, there are two modes of operation: StandAlone (SA) and Non-SandAlone (NSA). The SA mode leverages the 5GC together with a Next Generation Radio Access Network (NG-RAN) to provide end-to-end connectivity to users, while the NSA mode relies on the Evolved Packet Core (EPC) or 4G/LTE infrastructure to provide connectivity and uses the 5GC and RAN for additional functionality and improved performance. Both modes offer various features and benefits, and the choice of which mode to use depends on the specific requirements and goals of the deployment.

In SA mode, the other major 5G component besides the 5GC is the NG-RAN. The NG-RAN is the network subsystem that provides wireless connection between the UE and the 5GC. The NG-RAN is responsible for transmitting and receiving radio signals to and from the UE and to perform radio resource management tasks such as allocation of radio resources and interference management. The main entity of the NG-RAN is the gNodeB (gNB), which is the radio transmitter responsible for transmitting and receiving radio signals to and from the UE. The gNB may be further divided into a gNB-Central Unit (gNB-CU) and one or more gNB-Distributed Units (gNB-DUs) linked by the F1 interface. The F1 interface supports signalling exchange and data transmission between the endpoints, separates Radio Network Layer and Transport Network Layer, and enables the exchange of UE-associated and non-UE-associated signalling.

The NG-RAN is connected to the 5GC via the NG interface, and it uses the 5GC to control the communication between the UE and the data network. The NG-RAN and the 5GC work together to provide various services to the UE, including voice and data communication, location services, and access to the internet and other data networks. Finally, the Xn control plane (Xn-C) interface is defined between two NG-RAN nodes. The transport network layer built on this interface is based on the Stream Control Transmission Protocol (SCTP) running on top of IP. **Figure 1** illustrates the connection between the NG-RAN and the 5GC.

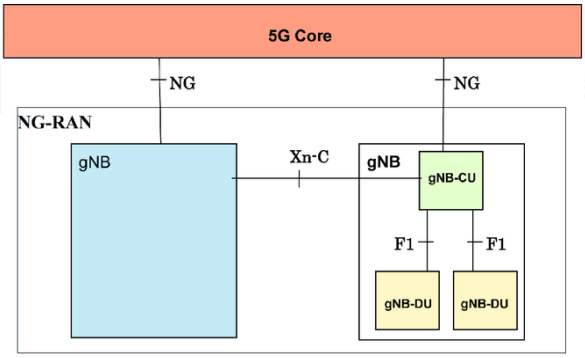


Figure 1: Architecture 5G base station gNB. Source: (Tikhvinskiy et al., 2020)

2.3 5G Verticals

A 5G vertical refers to a particular industry or market segment that a 5G network is designed to serve. Different industries may have specific requirements for connectivity or capabilities, and telecommunications companies may offer specialized solutions for each vertical to meet those needs. By segmenting the market into different verticals, companies can tailor their products and services to the specific needs of each industry. The main 5G verticals (5G-PPP JI, 2023) are:

- Automotive: Promoting self-driving vehicles and other transportation innovations, improving safety and reducing congestion.
- Manufacturing: Boosting the use of robotics and automation in manufacturing, improving efficiency and reducing costs.

- Media: Allowing higher-quality streaming of video and other media, as well as the use of virtual and augmented reality in the media industry.
- Energy: Facilitating the use of the Internet of Things (IoT) in the energy industry, allowing for the remote monitoring and control of power grids and other energy infrastructure.
- eHealth: Allowing remote surgery, telemedicine, and other healthcare services, permitting doctors to treat patients remotely and providing access to healthcare in underserved areas.
- Public Safety: Using faster response times for emergency services, as well as the use of drones and other technologies for search and rescue operations.
- Smart Cities: Enabling the use of IoT in city infrastructures, such as smart traffic systems and smart lighting, as well as the use of sensors and other technologies to improve services and living standards.

Overall, it is expected that 5G technology may enable the development of new and improved services across a wide range of vertical industries, with the potential to increase efficiency, improve safety, and enhance the quality of life for citizens.

2.4 5G Local Area Networks

A 5G Local Area Network (LAN) is a type of private cellular network that is designed specifically for use by enterprises. It integrates with an organization's existing Information Technology (IT) infrastructure to provide high-speed, predictable wireless connectivity with deterministic performance and latency for mission-critical digital initiatives across the enterprise. 5G LANs differ from commercial 5G services and Managed Service Providers (MSPs) in that they offer more control and flexibility to the enterprise, as well as the ability to integrate seamlessly with the organization's existing IT infrastructure and policies. 5G LANs use private spectrum and plug-and-play hardware and can be easily scaled and budgeted for without the need for specialized cellular network expertise. They are suitable for use in a variety of environments, including hospitals, factories, warehouses, and retail stores (Team Celona, 2020). Here are a few examples of how they might be used:

- Manufacturing: enabling IoT devices for instantaneous performance, output, and maintenance updates, and providing customized coverage and controls for applications, sensors, and devices.
- IIoT: supporting thousands of devices and sensors throughout industrial systems and offering wired-like reliability in complex environments with customized hardware, increased power levels, and multiple frequency bands.
- Healthcare: providing secure and reliable connections for patient room technologies and staff, and to configure service level objectives to serve patient health sensors, ventilators, clinical applications, and medical inventory levels.

Overall, 5G LANs offer a range of benefits for businesses and organizations that need high-speed, reliable wireless connectivity within a specific location. These networks can be deployed as standalone systems or as part of a larger 5G network, depending on the specific needs and requirements of the user.

2.5 5G Testbeds for IIoT integration and security research

There are recently developed testbeds for 5G research, as it is the case for the 5Greplay testbed (Salazar et al., 2021, which was designed for 5G security research and development of fuzzing platform for attack injection. BlueArch (Ghosh et al., 2019) uses several open-source components to provide a 5G customizable platform for experimentation of several scenarios, including IoT-based ones. (Amponis et al., 2022) also describes a testbed that shares some similarities with the one presented in this paper, designed as a cyber-range to test attacks against the 5G Core and RAN.

The testbed hereby presented distinguishes itself from previous efforts due to the fact that it's based on a cloud native deployment, also being designed to incorporate a mix of real and emulated equipment, supporting an IIoT scenario that can be used for research purposes in a protected environment. It provides a realistic reproduction of the characteristic 5G service footprint, also including user equipment and an underlying infrastructure very similar the one found on commercial hosting platforms.

3. Enabling Technologies

This section introduces the key technologies that enable/support the testbed. These technologies have been carefully selected based on their common integration capabilities, as well as their popularity and proven track record in the industry. Additionally, other important factors were taken into consideration, such as their stability,

security, scalability, ease of use, cost-effectiveness, and community support. Furthermore, the open-source nature of many of these technologies gives the advantage of being able to customize the technology to our specific needs, which can lead to cost savings and reduction of vendor lock-in. They also have a robust ecosystem of third-party tools and services, making them a suitable choice for organizations that want to integrate with their existing infrastructure.

3.1 Rancher's RKE and Server

Rancher Kubernetes Engine 2 (RKE2) (SUSE Rancher, 2023) is an open-source distribution for deploying and managing Kubernetes (K8s) (CNCF, 2022) clusters. It was chosen due to its ease of use, as well as for its opensource nature allowing for freedom in terms of usage and modification, providing flexibility in customization. Additionally, its reliability and robustness have been widely assessed in the scope of production environments and real-world scenarios.

The RKE2 deployment is complemented by a Rancher Server (SUSE Rancher, 2022) that provides a graphical user interface for managing a Kubernetes cluster, its containers, and applications. It is designed to make it easier for users to interact with their clusters and deploy and manage their applications. Some of the key benefits of using the Rancher server include its ease of use, which provides a user-friendly interface that makes it simple to manage and monitor a Kubernetes cluster, even for users who are not familiar with Kubernetes. Additionally, the Rancher server includes several functionalities that allow users to monitor and manage the resources within their cluster, such as the ability to view metrics, logs, and events for individual pods, as well as the ability to manage and scale deployments.

3.2 Open5GS and UERANSIM

Open5GS (Lee, 2022) is an open-source implementation of the 5GC and EPC written in the C programming language. It can act as a core network for New Radio (NR)/Long Term Evolution (LTE) setups and can operate in both StandAlone (SA) and Non-Standalone (NSA) modes. UERANSIM (Güngör, 2022) is the next generation of open-source 5G User Equipment (UE) and Next Generation Radio Access Network (NG-RAN) (gNodeB) implementation. It can be considered as a 5G mobile phone and a base station in basic terms, both running as virtualized instances. The UERANSIM project can be used to test the 5G main components and study the 5G system.

3.3 OpenEBS

OpenEBS is a technology that allows dynamic storage of persistent volumes (OpenEBS, 2021), making it a good fit for organizations with varying storage needs. It creates persistent volumes which are important for applications that need to retain data across restarts or failures. The automatic linking of a created volume with a specified application, such as a Database (DB), simplifies the process of setting up and configuring storage.

3.4 Kube-vip

Kube-Vip (Linux Foundation, 2022) provides Kubernetes clusters with a virtual IP and load balance for the control plane (to build a highly available cluster) and *LoadBalancer* type Kubernetes Services without relying on any external hardware or software. Regarding the cluster servers, there is a leader, and if the leader goes down, Kube-VIP quickly assigns another leader (about 10s) to always guarantee high availability.

3.5 OpenPLC

OpenPLC (OpenPLC Project, 2022) is the first fully functional standardized open-source PLC ("PLC", 2023), both in software and hardware. It is mainly used in industrial and home automation, IoT and SCADA research, as it carries built-in security features. It also does not require any additional licensing fees and allows for frequent updates and patches. The OpenPLC Project consists of two parts: Runtime and Editor. A Runtime is a portable software designed to run from the smallest of all microcontrollers (Arduino-compatible) to powerful servers in the cloud. The OpenPLC Editor is the software that runs on your computer and is used to create PLC programs. It is very simple to use and supports all languages that are compatible with PLC.

OpenPLC instances communicate with other PLCs primarily using Modbus (Modbus Organization, 2023) that runs over port 502 using Transmission Control Protocol (TCP)/Internet Protocol (IP), albeit other protocols, such as DNP3 (DNP ORG, 2023), are also supported. Modbus adopts a master-server communication pattern based on polled operations, where the slave synchronizes with the master by reading and writing specific registers that are configured by the user, such as output registers, input registers, and other types of registers. This allows the PLCs to exchange data and coordinate their actions to control and monitor industrial processes.

4. Design and Implementation of the Testbed

In this section, we will delve into the testbed from a user perspective, examining its structure and network topology. Additionally, we will cover how to check connectivity between PLCs and present a use case to illustrate and validate the testbed's capabilities.

4.1 User Viewpoint and Structure

The foundations of this testbed are based on a Kubernetes cluster, which enables the automatic management of resources according to predefined rules. The Rancher Kubernetes Engine 2 (RKE2) is the technology used for the production grade Kubernetes cluster distribution. A Rancher Server, complete with a graphical interface, was deployed within the cluster to control it. Kube-VIP was also deployed within the cluster to provide high availability using a floating virtual IP. To provide dynamic persistent volumes for the MongoDB database used in the 5GC, OpenEBS was deployed, which automates this process.

For the 5GC, the Open5GS implementation was used, which is composed of a set of functions deployed within the cluster. For users to connect to the 5G network, the radio simulator and user equipment are provided by UERANSIM, which is also deployed within the cluster. This implementation was used to provide access to the 5G network because it is inserted within the same namespace as the cluster, as otherwise it would not function properly. The implementation of the use case will be carried out by utilizing both OpenPLC and UERANSIM. In **Figure 2**, we can see a visual representation of the building blocks and tools used in the proposed testbed.

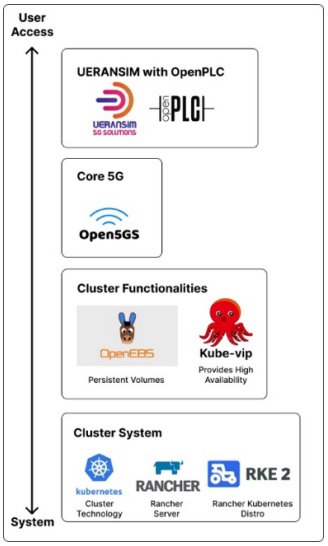


Figure 2: Diagram of every technology used from a user perspective access to the testbed.

The testbed specific structure is depicted in **Figure 3**. There are two ways to access the cluster: through a Virtual Machine (VM) that has a master node, or by connecting to an endpoint on the rancher server. Both methods require a Virtual Private Network (VPN) to the data center, managed by a hypervisor that oversees the allocation of resources, VMs, and networks, where this testbed is deployed. Additionally, it is possible to access one of the pods if they have an exposed web interface. The master nodes provide high availability using a floating Internet Protocol (IP) address. The Kube-VIP service allows for the creation of multi-node or multi-pod clusters, which can be used to provide even greater availability. In particular, the cluster can be configured to use an Address Resolution Protocol (ARP) mode, in which a leader is elected and inherits the virtual IP. This leader is responsible for load balancing within the cluster, allowing for the seamless integration of new worker nodes without interrupting the overall functionality of the cluster, even if the leader goes down.

The cluster operates by sharing resources between worker nodes, which Kubernetes uses to manage pods containing various technologies, including UERANSIM, OpenPLC, Open5GS functions, OpenEBS, Rancher Server, and Kube-VIP. These technologies are typically deployed in individual pods except for UERANSIM and OpenPLC, which are deployed together. These pods can be easily replicated to meet changing needs.

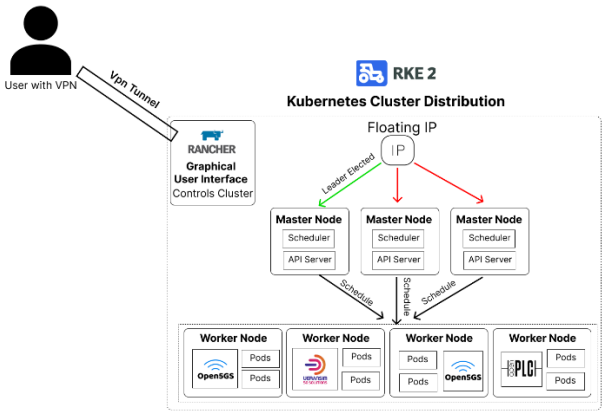


Figure 3: Structure diagram of the testbed.

Next, we will look at how the endpoints work in the Kubernetes cluster (**Figure 4**). In a K8s cluster, services typically have a fixed IP address that can be accessed over the internet without the need for an ingress. These services are configured with the "Load Balancing" type, which is provided by Kube-VIP and distributes the load evenly across all pods. Pods are one or multiple containers in runtime. The **Figure 4** shows 3 services with 3 pods each. For each service, there are 3 replicas of a pod. This example can be scaled as needed.

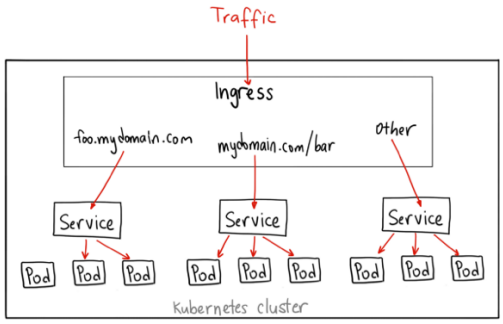


Figure 4: Diagram of services in detail. Source: (Abualrob, 2022)

4.2 Network Topology

For the network, we use port **group 223**, because it's configured with a DHCP Server and a VPN server facilitating the setup of the cluster and accessing it from a VPN client. **Figure 5** illustrates the testbed network topology.

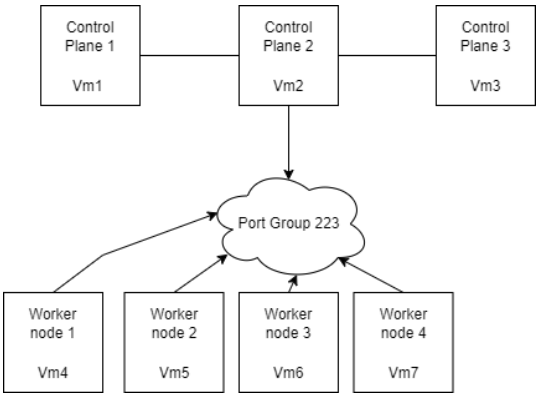


Figure 5: Network topology of our testbed.

The testbed implemented includes 7 VMs, three of them corresponding to the master nodes and having fixed IP addresses. These are responsible for managing the cluster, and resources and assigning tasks to the agent nodes (workers). Agent nodes(workers) have dynamic IP because they don't need to have static IPs, as they can be dynamically scaled up or down. These connect to the cluster from the virtual IP that makes the master nodes in high availability. There is a fixed IP range where the services are hosted automatically.

5. Use Case – OpenPLC communications using a 5G Core

This section describes a simple but representative Use Case: OpenPLC communications using a 5G Core. This use case demonstrates the capabilities of the testbed in the field of IIoT by creating a practical scenario where OpenPLC software is utilized within virtual containers and user equipment is used for network interfaces. The scenario involves connecting a master device to a slave device and verifying successful communication via the 5G network using the Modbus protocol. To test synchronization between OpenPLC registers through the 5GC, tools such as *tcpdump* (Tcpdump Group , 2023) and *Modbus poll (*Witte Software, 2023) can be used. *tcpdump* provides information on traffic between IP addresses and can verify the use of IPs belonging to the 5GC. *Modbus poll* can be used to read and write to registers belonging to the master and slave devices. **Figure 6** illustrates the communication between the OpenPLC instances.

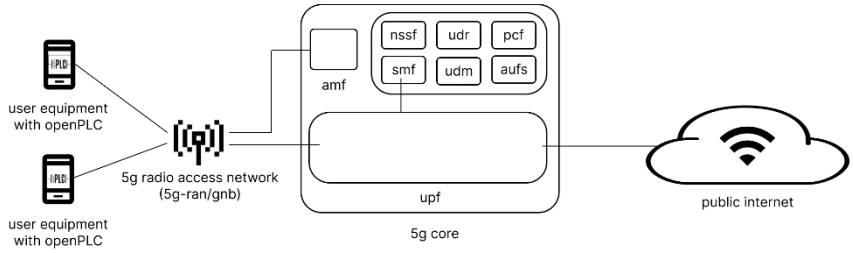


Figure 6: OpenPLC use case. The master device connects to the User Plane Function (UPF) in the 5GC, then reaches the slave device.

This IIoT scenario (see Figure 7) reproduces a hub-and-spoke topology similar to the one used in (Sousa et al., 2021), reproducing a topology with several geographically distributed elements. PLC master nodes query other PLC nodes to gather information from sensors, which are stored in specific Modbus holding registers – this scenario involves both horizontal (PLC-PLC) and vertical (PLC-HMI) communication patterns with polling cycles scheduled at 100ms intervals. Each Master has several configured PLC slaves, which are regularly pooled – the latter implement the control loops for locally attached sensors, also performing actuation. The underlying logic considers a process for controlling the water level control in a pumping station, using a level sensor and a water pump which is activated when a certain threshold is reached.

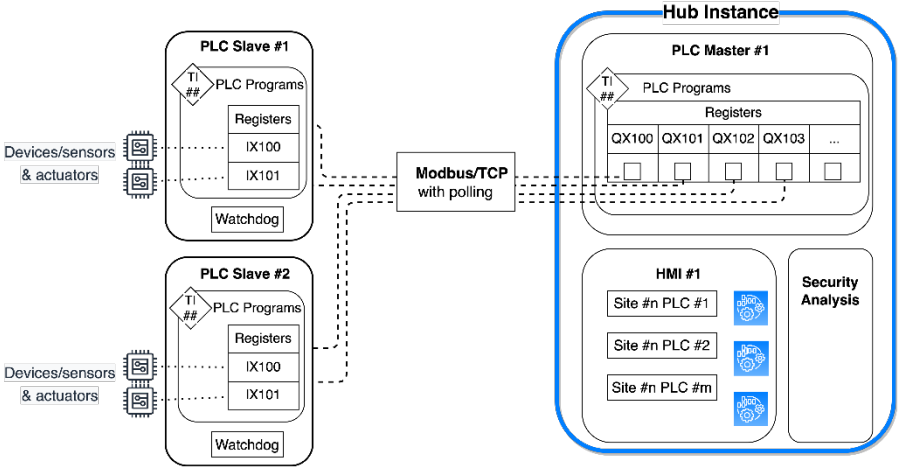


Figure 7: Hub-and-spoke architecture for the IIoT scenario

PLCs are programmed using Ladder Logic language, with the device and actuator information being mapped into PLC registers whose type is chosen accordingly with the nature of the involved data (Modbus is a very simple protocol with rudimentary data models) – for instance, discrete input coils (%IX addresses) can represent binary states (connected - 1/true, and disconnected - 0/false), while analog registers (%IW for input or %QW for outputs) can store values varying along a scale or range, being mapped in 16 bit registers. This architecture also enables the aggregation of data in Human-Machine Interface (HMI) consoles, providing informational dashboards can be designed to convey real-time information and also be used to trigger actions.

This architecture can accommodate a mix of real and emulated equipment, also being able to be fed by real data coming from a cyber-physical process used in production. In this latter case, it means that the proposed IIoT scenario can evolve into a Digital Twin, taking advantage of the specific 5G service assurance characteristics to offer an off-premises safety and security-enhancing service.

6. Conclusions

This paper has presented the development and implementation of a testbed for the evaluation of 5G and IIoT technologies. The testbed is based on open-source technologies such as UERANSIM as the Next Generation Radio Access Network (NG-RAN), Open5GS as the 5GC, OpenPLC to simulate PLCs, and RKE2 as the K8s distribution, which were chosen for their integration, popularity, and reliability. The testbed has been designed to provide a flexible and cost-effective solution for researchers and developers to simulate and benchmark their solutions in an easy-to-use framework.

To confirm the effectiveness of the testbed, a IIoT use case for OpenPLC communications using a 5G Core was implemented and the results showed successful communication between OpenPLC instances utilizing the 5GC. This use case not only validated the capabilities of the testbed in the field of IIoT but also demonstrated the potential for future advancements in industrial communication using 5G technology.

Acknowledgments

This work was supported in part by the Fundo Europeu de Desenvolvimento Regional (FEDER)-Competitiveness and Internationalization Operational Program (COMPETE 2020), Portugal 2020 Framework, in the scope of the Smart5Grid Project, under Grant POCI-01-0247- FEDER-047226; and in part by the FCT–Foundation for Science and Technology, Instituto Público (I.P.)/MCTES through National Funds, within the scope of Centre for Informatics and Systems of the University of Coimbra (CISUC) Research and Development Unit, under Grant UIDB/00326/2020 and Project UIDP/00326/2020.

References

3GPP TSG Group. (2019). 3GPP, Release 15. Available at: https://www.3gpp.org/specifications-technologies/releases/release-15

3GPP TSG Group. (2020). 3GPP, Release 16. Available at: https://www.3gpp.org/specifications-technologies/releases/release-16

3GPP TSG Group. (2022). 3GPP, Release 17. Available at: https://www.3gpp.org/specifications-technologies/releases/release-17

3GPP. (2022). 3GPP, a 5G analysis. Available at: https://www.3gpp.org/technologies/5g-system-overview

3GPP. (2022). 3GPP, about us. Available at: https://www.3gpp.org/about-us

5G-Infrastructure-PPP Joint Initiative. (2023). 5G And Verticals. Available at: https://5g-ppp.eu/verticals/

Abualrob, S. (2022). Kubernetes: Expose pod externally bypass service load balancing. Medium. Available at: https://medium.com/@suleimanabualrob/kubernetes-expose-pod-externally-bypass-service-load-balancing-bf89038afee2

Amponis, G. and Radoglou, P. and Thomas, G. and Ouzounidis, S. and Zevgara, M. and Moscholios, I. and Goudos, S. and Sarigiannidis, P. (2022) “Towards Securing Next-Generation Networks: Attacking 5G Core/RAN Testbed” in Proceedings of PACET 2022: PAnhellenic Conference on Electronics and Telecommunications, December 2-3, 2022, Tripolis, Greece.

CNCF (2022). Kubernetes Documentation. Available at: https://kubernetes.io/docs/home/

DAEnotes (2023). PLC Diagram. Available at: https://www.daenotes.com/electronics/industrial-electronics/PLC-programable-logic-control

DNP ORG. (2023). Overview Of DNP3 Protocol. Available at: https://www.dnp.org/About/Overview-of-DNP3-Protocol

Güngör, A. (2022). GitHub UERANSIM. Available at: https://github.com/aligungr/UERANSIM

In Wikipedia. (2023). PLC. Available at: https://en.wikipedia.org/wiki/Programmable_logic_controller

ITU. (2015). IMT-2020. Available at: https://www.itu.int/en/ITU-T/focusgroups/imt-2020/Pages/default.aspx

Lee, S. (2022). Open5gs - documentation. Available at: https://open5gs.org/open5gs/docs/
Linux Foundation. (2022). Kube-vip. Available at: https://kube-vip.io/docs/
Modbus Organization. (2023). The Modbus Official Site. Available at: https://modbus.org/
OpenEBS Project. (2021). OpenEBS. OpenEBS documentation: Openebs docs. Available at: https://openebs.io/docs/
OpenPLC Project. (2022). Open-source PLC software. Available at: https://openplcproject.com/
Pathak, R. (2022). Deploying 5g core network with open5gs and ueransim. Rahasak Labs. Available at: https://medium.com/rahasak/5g-core-network-setup-with-open5gs-and-ueransim-cd0e77025fd7
POWER Project. (2020). POWER - Empowering a digital future. Available at: https://www.cisuc.uc.pt/en/projects/power
Smart5Grid Project. (2021). Smart5Grid - Automated 5G Networks and Services for Smart Grids. Available at: https://www.cisuc.uc.pt/en/projects/smart5grid-automated-5g-networks-and-services-for-smart-grids
Ghosh, S. and Ugwuanyi, E. and Dagiuklas, T. and Iqbal, M. (2019) "BlueArch–An Implementation of 5G Testbed," Journal of Communications vol. 14, no. 12, pp. 1110-1118, 2019. DOI.: 10.12720/jcm.14.12.1110-1118
Salazar, Z. and Nghia, H. and Mallouli, W. and Cavalli, A. and Montes de Oca, E. (2021) 5Greplay: a 5G Network Traffic Fuzzer - Application to Attack Injection. In Proceedings of the 16th International Conference on Availability, Reliability and Security (ARES 21). Association for Computing Machinery, New York, NY, USA, Article 106, 1–8. https://doi.org/10.1145/3465481.3470079
Sousa, B. and Arieiro, M. and Pereira, V. and Correia, J. and Lourenço, N. and Cruz, T. (2021) ELEGANT: Security of Critical Infrastructures With Digital Twins, in IEEE Access, vol. 9, pp. 107574-107588, 2021, doi: 10.1109/ACCESS.2021.3100708
SUSE Rancher. (2022). Rancher Server v2.6. Available at: https://rancher.com/docs/rancher/v2.6/en/
SUSE Rancher. (2023). Rancher's next-generation kubernetes distribution. Available at: https://docs.rke2.io/
Tcpdump Group. (2023). TCPDUMP Official Website. Available at: https://www.tcpdump.org/
Team Celona. (2020). 5G LAN: What Is It & Why Will It Matter? Available at: https://www.celona.io/5g-lan/5g-lan
Tikhvinskiy, V. and Koval, V. (2020). Architecture 5G base station gNB. Available at: https://www.researchgate.net/figure/Architecture-5G-base-station-gNB_fig5_339059845
Witte Software. (2023). Modbus Poll Official Website. Available at: https://www.modbustools.com/modbus_poll.html

Cognitive Security: Facing Cognitive Operations in Hybrid Warfare[1]

Didier Danet
Saint-Cyr Military Academy, Guer, France

didier.danet@st-cyr.terre-net.defense.gouv.fr

Abstract: The digital space is now an active area of conflict. Attacks take many forms, to the point where concepts multiply and overlap. One concept in particular raises important questions: cognitive warfare. Cognitive warfare is an issue of concern to all countries, but there are very significant differences in approach between countries such as the United States or France and countries such as Russia, Iran or China. We try to show that these very different approaches leave Western countries unprepared for a global threat and poorly identified vulnerabilities. The concept of cognitive security therefore appears to be a promising avenue for reflection.

Keywords: Cyberdefense, Disinformation, Cognitive Warfare, Decision-Making Systems, Cognitive Security

1. Introduction

The digital space is an area of confrontation and conflict: cyber attacks, disinformation campaigns, espionage... In a world that is now “phygital”, distinguishing between the kinetic and non-kinetic dimensions of confrontations becomes more and more difficult, generating a certain confusion in concepts and categories. Hybrid war, electronic warfare, cyberwar, information warfare or cognitive warfare tend to overlap. This presentation is devoted to the concept of cognitive security, the interest of which is to take an overview of the operations carried out by certain countries such as Russia in intangible fields, to assess the threats and vulnerabilities and to propose appropriate responses.

2. From Information Operations to Cognitive Security: Old and New Threats

Cognitive warfare appears to be a concept still under construction in the various military doctrines of Western countries and their competitors. Amidst the diversity of interpretations that are given, a major difference between two main approaches is to be noted. Western countries emphasize the techno-scientific dimension and willingly link the notion of cognitive warfare to progress in bio-technologies and neuro-sciences. Countries like Russia have a more traditional approach. There, cognitive warfare is part of informational warfare.

2.1 The Western Approach to Cognitive Warfare: Technology First

In the West, many doctrinal documents pay attention to cognitive warfare. A common feature of this thinking is the tendency to favour a techno-scientific approach and to focus on the “weaponisation of biotechnologies and neuroscience”. (Giordano, 2017; Rickli and Ienca, 2021)

Thus, the French Minister of Defence mentioned “a new form of threat combining capabilities in information manipulation, disinformation, cybernetics, psychology, social engineering, biotechnology, and which could be summarised as a kind of new form of warfare: cognitive warfare. Cognitive warfare is the ability to exploit the vulnerabilities of the human brain using all the methods I have just mentioned.”

'This position is not isolated. Hacking the human brain is also a central element in the typology proposed by Hung and Hung. For these two authors, cognitive warfare and other types of warfare (information warfare, cyberwar…) “might contain the element of operations to influence and impact human cognition; however, only cognitive warfare is specifically dedicated to the control of the brain by incorporating militarised neuroscience into various practices”.(Hung and Hung, 2022)

2.2 The Russian Approach to Cognitive Warfare: Disinformation First

Originally, the major issues of cyber defense have revolved around attacks aimed at destroying or damaging vital military and industrial systems: administration, energy, transport, banks ... The attacks against Estonia (2007) or Georgia (2008) have attracted a lot of attention, and rightly so, since these attacks are still on the agenda with the invasion of Crimea (2015) and the Donbass (2022). The notion of "Cyber Pearl Harbor" reflects the fear of a

[1] The author thanks anonymous reviewers for providing helpful comments on earlier draft of the manuscript.

massive attack that could seriously disrupt or paralyze the entire life of a country or a region.(Dunn Cavelty, 2013; Lawson and Middleton, 2019)

But cyber-attacks such as the one against the French television channel TV5 Monde have shown that Russia can pursue another objective: to damage or destroy the image, reputation or credibility of its adversaries. (Brangetto and Veenendaal, 2016)

The "Doctrine of Information Security" formulates the ambition to guarantee Russia's sovereignty in cyberspace by providing "reliable information" on the policies of the country and its main leaders and by exercising digital influence. (Darczewska, 2015; Nocetti, 2015) This influence strategy mobilises all the media available in the digital space and ensures the unity and consistency of their discourse.

Historically, Russia's disinformation operations were first deployed in Europe.(Limonier and Audinet, 2017) Then, for a variety of reasons, they expanded to Southeast Asia and Africa.(Limonier, 2018)

In Europe, Russian disinformation operations involve the traditional media, particularly television channels, which have benefited from a sharp increase in budgets and resources since the beginning of the 2010s. (Fletcher *et al.*, 2018) Before the invasion of Ukraine and the sanctions imposed on RT and Sputnik, half of its viewers were in ten European countries including France, Germany and the United Kingdom.. In addition to the influence of traditional media on the Internet, social media are also very much in demand. The two dimensions are totally linked, as RT and Sputnik are very present on all applications such as Facebook, Twitter, Youtube, Instagram...

Since the mid-2010s, the cost-effectiveness of Russian disinformation operations in Europe has decreased, in particular thanks to the efforts undertaken by the European Union and its member countries to denounce these operations. Since the invasion of Ukraine, the two main vectors of Russian influence, RT and Sputnik, have been shut down, forcing Russia to change its focus.

In this respect, the African continent is a new priority. Until the end of the 2010s, Africa occupied a marginal place in Russian foreign policy, which was reflected in the very limited presence of cultural and media relays on the continent. In a very short period of time, this presence has increased dramatically, in particular as a result of an adaptation to the local context: a backbone based on two media (RT and Sputnik) and alternative and conspiracy-oriented content.

The adaptation to the local context is particularly impressive in terms of the construction of networks that promote the Russian vision.(Clifford and Gruzd, 2022) This adaptation also involves the diversification of the actors who carry Russian interests in the region. Alongside state media such as RT and Sputnik, private disinformation entrepreneurs are playing a very significant role. (Limonier and Laruelle, 2021) The result of this spectacular progression of Russian vectors of influence in Africa is a full-scale attack on the Western presence (especially the French) on the continent and, in particular, on the Barkhane military operation.(Douzet *et al.*, 2021)

In conclusion, it appears that the real threat facing Westerners in the field of cognitive warfare is not related to research programs aimed at 'hacking the brains' of individuals through biotechnology and neuroscience processes. Russia's actions in the digital space are part of a very classic policy of influence, but carried out with significant means, a global approach, a constant strategy and adaptation to the local context. (Strategic Communications, Task Forces and *et al.*, 2023) The real threat is therefore the ability of Russian propaganda to directly or indirectly influence Western decision-making processes and to challenge the image, reputation or credibility of the latter. This is where the issue of cognitive security comes in, given the vulnerabilities within Western decision-making systems.

3. Cognitive Threats: A Model

Russia's disinformation campaigns have raised awareness of the danger represented by influence operations aimed at destabilising institutions (TV5 Monde) or democratic processes (French presidential elections in 2017). The European Union and its member countries have reacted in several ways: ad hoc legislation or the creation of bodies to combat fake news such as the EU Disinformation Lab With the closure of RT and Sputnik following the invasion of Ukraine, it can be said that the threat has significantly diminished on the old continent. But we have seen that it has in fact moved to areas where our interests are now under pressure from influence campaigns, notably in Africa where France's positions are in difficulty with the imposed withdrawal of its armed forces from the Central African Republic, Mali and Burkina Faso. Above all, a new threat is emerging with the development of campaigns aimed at destabilising Western decision-making systems.

To highlight the nature and modalities of these "cognitive" campaigns, a simplified model of a collective decision-making process is proposed. It is an Agent / Environment / Decision model. The agent is a civil or military organisation which has a collective decision system integrating an expert/decision maker pair. This agent is given objectives. It observes the environment and takes decisions likely to lead to the objective. The decision process is of the OODA type.

To highlight the nature and modalities of these "cognitive" campaigns, a simplified model of a collective decision-making process is proposed. It is an Agent / Environment / Decision model. The agent is a civil or military organisation which has a collective decision system integrating an expert/decision maker pair. This agent is given objectives. It observes the environment and takes decisions likely to lead to the objective. The decision process is of the OODA type.

In the light of this model, three types of cognitive threats can be identified.

The first possible action concerns the environment in which the organisation must act and make decisions. As Dave Snowden has shown, a decision process of the type Sense / Analyse / Decide / Act (i.e. the OODA loop mechanism) is well suited to so-called complicated environments but where the universe remains ordered. (Snowden and Boone, 2007)

However, OODA-type decision-making processes can be destabilised in disordered environments, which Dave Snowden calls complex (cause and effect relationships seem to have disappeared even if patterns gradually emerge) or chaotic (no order can be discerned in an environment in crisis). The role of management teams changes profoundly and the appropriate leadership style is based on specific qualities such as the ability to take responsibility, to think outside the box and to bring solutions to the surface rather than to arbitrate. It is inferred that a possible attack on the decision-making systems of Western organisations, which are mainly based on OODA-type schemes, is to cause confusion or chaos in the environment. This can be interpreted as Russian propaganda in Africa manipulating public opinion in Mali or Burkina Faso, provoking demonstrations to block military convoys and placing those in charge of the convoys in situations that could degenerate very quickly.

The second type of cognitive attack is aimed at disrupting the critical cognitive infrastructure that allows deployed forces to receive the information they need, communicate with each other and maintain a constant link with headquarters. These attacks can be brutal, for example, the destruction of a communication system. They can be more insidious: flooding a communication channel to dilute relevant data in a sea of noise or altering the data flowing through the infrastructure.

This last type of attack directly targets the decision-making systems of Western organisations. The aim is to take advantage of existing vulnerabilities in these systems to make them commit significant errors that will prevent them from carrying out their missions or discredit them in the eyes of the population.

The next section will be devoted to an analysis of the possible vulnerabilities of Western decision-making systems.

4. Westen Organisations at Risk: Vulnerabilities

Faced with the three types of attack that the Agent/Environment/Decision model highlights, why might Western civilian or military organisations find themselves in trouble? The answer is well known to social scientists who are interested in the structures and processes of decision-making in organisations (management sciences, organisational sociology, behavioural psychology). There are three main reasons why collective decision-making systems go wrong. (Morel, 2002).

4.1 Cognitive Errors

The first source of error can be described as a "cognitive" error. The organisation does not have the necessary input data at the time it makes its decision. It does not have the right theories, concepts or tools to correctly understand the situation and the possible solutions. It engages in "cognitive tinkering" characterised by shortcuts in thinking, approximations in orders of magnitude, erroneous representations, etc. We have seen in relation to attacks on critical cognitive infrastructure that this can tend to introduce falsified data into decision-making systems. The latter then applies the analysis schemes and decision mechanisms correctly but, working from a representation of the environment that does not conform to reality, the decisions adopted are not adapted to the situation. Resources are wasted, the results obtained are not in line with the objectives, the organisation's action is mocked and its credit in the public opinion is degraded.

This type of cognitive attack is the most frequent and oldest mode. It is the one that the Russian eco-system has been engaging in in Africa since the mid-2010s.

The second type of cognitive error does not involve any particular manoeuvre by the adversary. It is more simply a "defeat of thought". The organisation that is given a mission does not make the effort to understand the terrain, is satisfied with simplifying reading grids, a supposed knowledge based on experience and, in short, a damaging intellectual laziness.

Finally, the representation error consists of transposing a certain conception from a source domain where it may be relevant to a target domain where there is no guarantee that it will be relevant. This type of error is related to what Daniel Kahneman calls “cognitive tinkering”. (Kahneman, 2012) One could quickly summarise this point by saying that “cognitive tinkering” consists of an organisation substituting a mode of reasoning based on intuition for a mode of reasoning of an analytical nature. This bricolage can be explained by many reasons: the sense of urgency, the impression of having already encountered a comparable situation, the weight of prejudices, the strength of intuition...

4.2 Collective Errors

Collective errors are inherent in the structure and functioning of groups (experts / decision-makers) that have to make a decision. It is widely recognised that a group decision mobilises different skills and abilities, which enriches the thinking and decision-making process, generally producing more relevant decisions. However, it should not be ignored that the group does not always function well and that dysfunctions may appear which will be the cause of collective errors.

A first source of collective error comes from a poor structuring of the group and the distribution of roles within it. For example, excessive fragmentation of decision-making groups creates "gaps" which are sources of friction and poor information transmission The adversary has little ability to influence the structure of such groups. But the bureaucratic nature of Western organisations, especially when they are public or international, tends to generate this type of vulnerability.

The second type of collective error is related to the poor functioning of groups. This functioning can be affected by multiple difficulties. The group may be excessively inward-looking and lose contact with its environment. This risk exists, for example, when a Western organisation is projected into a theatre of crisis or war and locks itself up in fortified camps for understandable security reasons. However, this confinement can produce a form of watertightness that is damaging to local actors and the Western actors. Another example is the communication procedures within a decision-making group. We are thinking here of the phenomena of "groupthink" (Janis, 2008) or the excess of implicit communication or the refusal of procedures aimed at formalising exchanges within the group...

This is not to say that Western organisations are riddled with weaknesses in their structures and collective functioning. One could even argue that, given the performance of the Russian army in Ukraine, our adversaries are probably even more affected than we are. But there is no reason to believe that Western organisations are immune and not at risk of making collective mistakes. The problem is that they are not diagnosed and the risk can arise here and there without being anticipated.

4.3 The Collapse of Sense-Making

Finally, the most serious mistakes come from the collapse of sense-making. As Karl Weick has shown, the collapse of meaning means that a group loses sight of the goal and focuses on the means. (Weick, 1993) The priority that should be given to the "why" of the action disappears in favour of the "how" of acting in uncertainty. This can have dramatic consequences as the objective is lost sight of and the implementation of the group's technical skills can lead in a completely different direction from that desired.

In difficult environments, trust and cohesion are two crucial characteristics of teams that do not collapse. Trust and cohesion may be jeopardized because of targeted influence campaigns. These operations make extensive use of the presence of civilian or military personnel in digital space (“digital fingerprinting”). The issue of “digital fingerprinting” was first raised because of the information leaks that can result from carelessness on the part of personnel on mission. The case of Strava watches used by military personnel to record their sporting activities has been widely studied and commented upon. (Cattaruzza and Taillat, 2018; Mericskay, Noucher and Roche, 2018; Condorcet, 2020) Similarly, the impact of videos or messages posted on social networks is the subject of increasingly common best practice guides. These mistakes obviously have consequences, not least because of

the spectacular development of OSINT techniques and tools. (Limonier 1 and Audinet 2, 2022; Renault 1, Charon 2 and Laurençon 3, 2022)

In addition to the problem of data leakage, the presence of people in the digital space opens the way for targeted influence operations that aim to undermine the trust of individuals in institutions, in their leaders or in their peers. This is not to say that all social networks inevitably breed distrust and division. But some of them base their ability to attract and retain users on such processes. For example, Facebook was a privileged vector of Russian influence during the US (2016) or French (2017) presidential elections insofar as the application exacerbates social feedback loops and emotional reactions.(Figeac *et al.*, 2019)

In conclusion, it appears that Western organisations deployed in crisis or conflict areas are weakened by the vulnerability of their collective decision-making processes. These vulnerabilities are characteristics common to all organisations. However, not all organisations have the same sources of vulnerability or the same degree of fragility due to these vulnerabilities. The problem is not to stigmatise the existence of these sources of vulnerability and the ambition should not be to eliminate them entirely. On the other hand, it is quite dangerous not to have a comprehensively designed and periodically applied assessment system. Ignorance of the current state of organisations, particularly those operating in difficult environments where the risks of Russian influence operations are known, does not allow for the anticipation of problems that may arise and does not allow for measures to be taken to strengthen “prone-crisis” organisations.

5. Enhancing Cognitive Security in Crisis Environments

Identifying threats related to Russian influence operations in both Europe and Africa, combined with the likely vulnerabilities of Western decision-making systems, leads to suggest measures to ensure the cognitive security of these systems. These recommendations can be classified according to the three types of attacks that have been underlined.

5.1 Cognitive safety in the face of environmental degradation

In disordered universes (complex or chaotic), the OODA loop (or Sense / Analyse / Decide / Act) is no more relevant and decision-makers must turn to processes where they must think outside the box, take the risk of acting without certainty and assume significant responsibilities in the event of failure. Three sets of measures are suggested to strengthen the cognitive security of the organisation.

First, anticipating adverse moves. Russian informational attacks follow relatively simple and constant patterns. (Chew and Turnley, 2017; Innes and Dawson, 2022; Olchowski, 2022) The propaganda arguments are few and relatively crude: the colonial past of the West, the hidden economic underpinnings of their interventions, the alternative reality and the conspiracies... (Nocetti, 2015) Similarly, the vectors of these operations are well identified: official Russian media and local partners, social networks of oligarchs and associated African entrepreneurs... It is therefore possible to anticipate informational attacks or to detect them at the earliest stages of their development.

Second, countering Russian narratives involves both refuting and responding to them. The actions to be taken are not always simple. In 2020, Facebook announced that it had closed 150 fake accounts, some linked to Russia and others to the French army. According to a Graphika report, the fake accounts linked to the French army were intended to respond to a Russian campaign in Mali by countering accusations of neo-colonialism with "posts" and cartoons focusing on the looting of mining resources by the mercenaries of the Wagner Group. (Graphika & the Stanford Internet Observatory, 2020) However, the ‘faux-nez’ used to carry out this retaliation were not sufficiently refined to escape Facebook's control procedures. Conversely, Western actions can be successful. For example, in April 2022, French forces in the Gossi region of Mali were able to film mercenaries from the Wagner Group staging a fake mass grave in order to accuse the French army. The dissemination of the images provided evidence of a staged operation aimed at turning the population against Operation Barkhane. (Robin, 2022)

Finally, if attacks result in significant environmental degradation, decision-makers must be prepared to engage in complexity or chaos management procedures. It is then necessary to ensure that all actors on the ground have these particular capabilities and, if these are not deemed complete enough, to set up training and education programs.

5.2 Cognitive security in the face of attacks on critical cognitive infrastructure

Attacks on critical cognitive infrastructure can be easily detected, especially when they aim to interrupt the attacked organisation's ability to perceive the environment and communicate. The destruction of a submarine cable or a terrestrial network is immediately perceived by users. However, if the attacker's aim is to drown the target in a flood of false information or noise, the intervention is more difficult to perceive. Therefore, three measures could be useful in this case.

The first is to systematically establish territorial connectivity diagnostics. These diagnoses aim to assess the digital information flow capacities in a given region, in particular the capacity that an operator such as Russia might have to divert internet flows to traffic nodes where it would be able to control and modify their content.

A second set of measures could aim at strengthening data-centric security measures. (Bayuk, 2009; Hennessy *et al.*, 2009; Arora, Raja and Bahl, 2018) These technical measures consist of ensuring that sensitive data is stored and circulates under the control of the organisation. For example, specific devices check that such data is not sent to unauthorised recipients or that it is present in certain areas of the information system. Such measures already exist in many organisations. They should be evaluated and strengthened if necessary.

Finally, the information systems security paradigm could be questioned as it seems to have reached certain limits. Traditionally, the protection of information systems has been seen as a fortress whose walls must be constantly raised to cope with the growing resources of attackers. However, the requirements of security designed in this way are becoming increasingly onerous and breaches due to the "human factor" are multiplying. Some authors advocate the development of a new approach in which the fortress is considered to be already invaded, thus forcing new behaviours on the part of the defenders of the system and its users. (Leuprecht, Skillicorn and Tait, 2016)

5.3 Cognitive security and the Collapse of Sense Making

Not all organisations are equal when faced with difficult, crisis or conflict situations. Some collapse ("prone-crisis organisations") (Mitroff *et al.*, 1989; Pauchant and Mitroff, 1992) while others prove surprisingly resilient (Highly Reliable Organisations). The parameters that explain this difference in response have been well identified in the academic literature. (Roberts, 1990) It is all the more surprising to note that audit mechanisms are not systematically put in place to identify and correct any weaknesses. The risks linked, for example, to an oversimplification of the grids for understanding situations or to an excessive polarisation of opinions within a unit should, in our opinion, be the subject of particular attention.

6. Conclusion

The term "cognitive warfare" has generated a certain amount of excitement, fuelled in particular by the spectacular progress of various scientific disciplines such as biotechnology, neuroscience and artificial intelligence. However, defining cognitive warfare by the famous expression "hacking the enemy's brain" amounts to focusing on a hypothetical threat, whereas our adversaries are practising a less sophisticated but equally dangerous "cognitive warfare", in particular because of the ignored vulnerabilities in our information and decision-making systems. Cognitive warfare should mainly be seen as an extension of information warfare. In this respect, threat analysis has made great progress, as demonstrated by the documents recently published by the European Union. (ENISA and EEAS, 2022; Strategic Communications, Task Forces and *et al.*, 2023) In particular, knowledge of the Russian disinformation eco-system, the analysis of the behaviour of actors or the study of speeches allows for early detection and a more appropriate response to disinformation manoeuvres. However, a grey area remains in this picture: the analysis of vulnerabilities that are exploited by attackers to achieve their objective. It is these organisational vulnerabilities that need to be better explored to address the threat of cognitive warfare.

References

Arora, A.S., Raja, L. and Bahl, B. (2018) 'Data centric security approach: A way to achieve security & privacy in cloud computing', in *Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)*, pp. 26–27.

Bayuk, J. (2009) 'Data-centric security', *Computer Fraud & Security*, 2009(3), pp. 7–11.

Brangetto, P. and Veenendaal, M.A. (2016) 'Influence cyber operations: The use of cyberattacks in support of influence operations', in *2016 8th International Conference on Cyber Conflict (CyCon)*. IEEE, pp. 113–126.

Cattaruzza, A. and Taillat, S. (2018) 'Les enjeux de la numérisation du champ de bataille', *Dynamiques internationales* [Preprint], (13).

Chew, P.A. and Turnley, J.G. (2017) 'Understanding Russian information operations using unsupervised multilingual topic modeling', in *Social, Cultural, and Behavioral Modeling: 10th International Conference, SBP-BRiMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings 10*. Springer, pp. 102–107.

Clifford, C. and Gruzd, S. (2022) 'Russian and African media: Exercising soft power', *South African Institute of International Affairs Policy Insights,(125)*, 1, p. 18.

Condorcet, A.C. (2020) 'Quelle place pour les cartes dans l'analyse politique de l'espace?'

Darczewska, J. (2015) *The devil is in the details. Information warfare in the light of Russia's military doctrine*. Warsaw: OSW Center for Eastern Studies (Point of View, 50).

Douzet, F. *et al.* (2021) 'Mapping the spread of Russian and Chinese contents on the French-speaking African web', *Journal of Cyber Policy*, 6(1), pp. 50–67.

Dunn Cavelty, M. (2013) 'Un Cyber Pearl Harbor : quelle probabilité à court terme ?', *Défense Sécurité Internationale (DSI)*, Numéro Spécial(32), pp. 30–32.

ENISA and EEAS (2022) *Foreign Information Manipulation Interference (FIMI) and Cybersecurity - Threat Landscape*, *ENISA*. Available at: https://www.enisa.europa.eu/publications/foreign-information-manipulation-interference-fimi-and-cybersecurity-threat-landscape (Accessed: 23 February 2023).

Figeac, J. *et al.* (2019) 'Facebook favorise-t-il la désinformation et la polarisation idéologique des opinions?', *Questions de communication*, (36), pp. 167–187.

Fletcher, R. *et al.* (2018) 'Measuring the reach of" fake news" and online disinformation in Europe', *Australasian Policing*, 10(2).

Giordano, J. (2017) 'Weaponizing the brain: Neuroscience advancements spark debate', *National Defense*, 6, pp. 17–19.

Graphika & the Stanford Internet Observatory (2020) *More-Troll Kombat*. Available at: https://graphika.com/reports/more-troll-kombat/ (Accessed: 10 February 2022).

Hennessy, S.D. *et al.* (2009) 'Data-centric security: Integrating data privacy and data security', *IBM Journal of Research and Development*, 53(2), pp. 2–1.

Hung, T.-C. and Hung, T.-W. (2022) 'How China's Cognitive Warfare Works: A Frontline Perspective of Taiwan's Anti-Disinformation Wars', *Journal of Global Security Studies*, 7(4), p. ogac016.

Innes, M. and Dawson, A. (2022) 'Erving Goffman on misinformation and information control: the conduct of contemporary Russian information operations', *Symbolic Interaction*, 45(4), pp. 517–540.

Janis, I.L. (2008) 'Groupthink', *IEEE Engineering Management Review*, 36(1), p. 36.

Kahneman, D. (2012) *Thinking Fast and Slow*. Available at: https://www.youtube.com/watch?v=GrFHRtGmE5I.

Lawson, S. and 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).

Leuprecht, C., Skillicorn, D.B. and Tait, V.E. (2016) 'Beyond the Castle Model of cyber-risk and cyber-security', *Government Information Quarterly*, 33(2), pp. 250–257.

Limonier 1, K. and Audinet 2, M. (2022) 'De l'enquête au terrain numérique: les apports de l'Osint à l'étude des phénomènes géopolitiques', *Hérodote*, (3), pp. 5–17.

Limonier, K. (2018) 'Diffusion de l'information russe en Afrique: essai de cartographie générale', *Note de recherche*, 66.

Limonier, K. and Audinet, M. (2017) 'La stratégie d'influence informationnelle et numérique de la Russie en Europe', *Hérodote*, (1), pp. 123–144.

Limonier, K. and Laruelle, M. (2021) 'Russia's African Toolkit: Digital Influence and Entrepreneurs of Influence', *Orbis*, 65(3), pp. 403–419.

Mericskay, B., Noucher, M. and Roche, S. (2018) 'Usages des traces numériques en géographie: potentiels heuristiques et enjeux de recherche', *L'information géographique*, (2), pp. 39–61.

Mitroff, I.I. *et al.* (1989) 'Do (some) organizations cause their own crises? The cultural profiles of crisis-prone vs. crisis-prepared organizations', *Industrial Crisis Quarterly*, 3(4), pp. 269–283.

Morel, C. (2002) 'Les décisions absurdes', *Paris, Gallimard*, 2012.

Nocetti, J. (2015) 'Guerre de l'information: le web russe dans le conflit en Ukraine', *Focus Strat*, 62, pp. 1–47.

Olchowski, J. (2022) 'How to Weaponise Information: Russian Patterns', in *Disinformation, Narratives and Memory Politics in Russia and Belarus*. Agnieszka Legucka&Robert Kupiecki. Routledge, pp. 59–71.

Pauchant, T.C. and Mitroff, I.I. (1992) *Transforming the crisis-prone organization: Preventing individual, organizational, and environmental tragedies*. Jossey-Bass.

Renault 1, C., Charon 2, P. and Laurençon 3, F. (2022) 'Renseigner autrement? Trajectoires de l'Osint dans les services de renseignement', *Hérodote*, (3), pp. 19–30.

Rickli, J.-M. and Ienca, M. (2021) 'The security and military implications of neurotechnology and artificial intelligence', *Clinical Neurotechnology Meets Artificial Intelligence: Philosophical, Ethical, Legal and Social Implications*, pp. 197–214.

Roberts, K.H. (1990) 'Some characteristics of one type of high reliability organization', *Organization Science*, 1(2), pp. 160–176.

Robin, M. and Kimberly (2022) *La doctrine cyber française, continuité et renforcements depuis 2008. La politique cyber française durant le mandat d'Emmanuel Macron (2017-2022).* Louvain: UCL, Faculté des sciences économiques.

Snowden, D.J. and Boone, M.E. (2007) 'A leader's framework for decision making', *Harvard business review*, 85(11), p. 68.

Strategic Communications, Task Forces and *et al.* (2023) *1st EEAS Report on Foreign Information Manipulation and Interference Threats | EEAS Website*. Brussels: European External Action Service. Available at: https://www.eeas.europa.eu/eeas/1st-eeas-report-foreign-information-manipulation-and-interference-threats_en (Accessed: 11 February 2023).

Weick, K.E. (1993) 'The collapse of sensemaking in organizations: The Mann Gulch disaster.', *Administrative science quarterly*, 38(4), pp. 628–652.

JTF-ARES as a Model of a Persistent, Joint Cyber Task Force

Charlie Donnelly[1] and Marcel Stolz[2]
[1]University of Chicago, Chicago IL, USA
[2]University of Oxford, Oxford, UK

csdonnelly@uchicago.edu
marcel.stolz@univ.ox.ac.uk

Abstract: Military involvement in cyberspace has traditionally been limited to operations in service of "kinetic," or physical, missions. Military cyberoperations are therefore usually described using traditional "kinetic" descriptors and rarely articulate cyber-related goals that are independent of kinetic operations. Recently, the concepts of "persistence" and "jointness" have been increasingly used by the U.S. Cyber Command to describe cyberoperations. Persistence describes operations that focus on a target over time (in contrast to the episodic "response" concepts articulated in kinetic warfare). "Jointness" describes working across group or agency lines. This paper will investigate the effectiveness of "persistent" and "joint" task forces in accomplishing cyber-related goals by means of a case study of Joint Task Force – ARES ("JTF-ARES"). JTF-ARES was set up as a task force by the U.S. Cyber Command to disrupt ISIS cyberoperations – a singularly cyber (as opposed to kinetic) goal. By contrasting the approach of JTF-ARES with the existing history of US operations in cyberspace, militaries can apply JTF-ARES' successful approach to accomplish future cyber-related goals that are independent of kinetic military units. After discussing a brief history of the U.S. Cyber Command and defining the terms "persistence" and "jointness," the paper discusses JTF-ARES' successful operation and contributing factors, most notably its organization within the U.S. Cyber Command. Next, it explores a counterfactual organization of JTF-ARES, suggesting that alternative organizational structures would likely have ended in failure and highlighting factors that may have influenced its success. Furthermore, the paper discusses the administrative challenges associated with creating a JTF, which include administration hurdles as well as collaboration and training requirements specific to joint operations. Since JTF-ARES deviates from traditional organizational structures within U.S. Cyber Command, this paper articulates criteria for creating a joint, persistent cyber task force, which militaries may find useful when considering how to implement cyber-specific goals. The first criterion concerns the operations required for the mission – namely, are reconnaissance, offensive, and defensive cyberoperations required? The second criterion asks whether the cyberoperation has a uniquely cyber-oriented end state: for missions with non-kinetic goals, it may be helpful to consider a joint, persistent task force.

Keywords: Military Cyberoperations, Task Force, Cyber Organization, Offensive Cyberoperations, U.S. Cyber Command, Kinetic Warfare

1. Introduction

The increasing relevance of cyberspace as an operational domain of the US military has necessitated dedicated government cyber units. These units are often created in service of military organizations, and therefore are conceptualized using military terminology pertaining to "kinetic," or physical, responses. In recent years, however, a "persistence" approach has been used to describe activities that are continuous in nature and may include several strategies, as opposed to episodic defensive or offensive cyberoperations. This new articulation of government cyberoperations is seen clearly in the creation of Joint Task Force – ARES, a subsection of the U.S. Cyber Command established to address ISIS operations. This paper will evaluate the effectiveness of "jointness" and "persistence" in achieving cyber goals that do not directly support kinetic operations. "Jointness" and "Persistence" are evaluated via a case study of JTF-ARES, and its organization in contrast to the rest of the U.S. Cyber Command.

Section 1 of this paper describes a brief history of the U.S. Cyber Command and traditional doctrines around cyberspace operations. Section 2 discusses the relatively recent terminology of "persistence" and jointness" used to describe cyberoperations. Section 3 discusses the case study of JTF-ARES, which can be considered a concrete example of persistent, joint warfare in cyberspace, and illustrates the potential of "single-purpose" task forces to accomplish ambitious cyber-related goals. Subsection A considers the creation of JTF-ARES as the first collaboration between the NSA and the U.S. Cyber Command, and subsection B describes the novelty of the task force in the history of USCYBERCOM operations. Section 4 explores counterfactual scenarios: namely, why JTF-ARES was a dedicated unit as opposed to an existing U.S. Cyber Command service component. Section 5 describes some of the administrative challenges associated with creating JTF-ARES, and it suggests questions to

determine the suitability of a JTF in future operations. Section 6 articulates two potential criteria for establishing future JTFs, and section 6.2 applies these criteria to discuss hypothetical situations in which a persistent JTF may be useful. Lastly, section 6.3 discusses areas of potential future research.

2. A Brief History and Evolution of the U.S. Cyber Command

The U.S. Cyber Command (USCYBERCOM) was established in 2010 by Secretary of Defence Robert Gates as a sub-unified command of U.S. Strategic Command following internal penetration testing and major data breaches from foreign adversaries. Before its establishment, individual cybercommand units were separated into defensive and offensive units, and later unified into USCYBERCOM in 2009 (Command History, n.d.). Along with the 2009 establishment of USCYBERCOM, service subunits were created to serve the Army, Air Force, Tenth Fleet, and Marine Corps (Command History, n.d.). In 2014, the Cyber National Mission Force (CNMF) and Joint Force Headquarters–DoD Information Network (JFHQ-DoDIN) were added. The CNMF was designed to "synchroniz[e] full-spectrum cyberspace operations to disrupt, degrade and defeat malicious cyber actors," and is described as "broad, continuous, joint, and enduring in nature" (Cyber National Mission Force Public Affairs, 2022). The CNMF is the military's newest sub-unified command, and may conduct offensive or defensive attacks.

The JFHQ-DoDIN was created in 2015 to defend DoD networks (Cyber National Mission Force Public Affairs, 2022). General Paul Nakasone serves as the leader of both the USCYBERCOM and the NSA (Inamete, 2022).

The history of the US military in cyberspace is relatively brief when compared with traditional "kinetic" operations, where "kinetic" is defined as "actions designed to produce…physical damage to…or destruction of targets" (Integration and Synchronization of Joint Fires, 2018). Cyberspace was only deemed a domain akin to air, land, sea, and space, in 2004, and as a result, cyberoperations are often conceptualized using kinetic terminology, which will be further discussed in later sections (Proctor, 2006).

3. Key Definitions: "Persistence," and "Jointness"

The words "persistence" and "jointness" have informed the U.S. Cyber Command's strategy under General Nakasone and are helpful descriptors for JTF-ARES. The doctrine of "persistence" is often described as "defending forward," which has been articulated in 2018 DoD strategy documents (Warner, 2020). "Defending forward" typically involves halting threats before they become offensive attacks, and continuously monitoring malicious cyberactivity.

The term "persistence" in cyberspace has traditionally been defined in opposition to a "response-force concept," or "holding forces in reserve for war or responding to attacks after the fact" (Schneider et al., 2020). Response-force is best described as an ad-hoc and episodic approach to warfare, which derives from traditional offensive and defensive separation that is characteristic of kinetic, or physical, warfare. General Nakasone (Commander of the U.S. Cyber Command and Director of the National Security Agency) characterizes cyberspace as a domain that requires reconnaissance, defense, and offense – the US must accordingly "operate continuously to seize and maintain the initiative in the face of persistent threats" (Joint Force Quarterly Issue 92, 2019). Nakasone also claims that USCYBERCOM was initially a response force, designed to defend against attacks – it wasn't until 2018 that the doctrine of a "persistent" force was finally articulated in USCYBERCOM's vision document (Joint Force Quarterly Issue 92, 2019). This shift in approach from "response" to "persistence" helps explain the context for the creation of JTF-ARES, which was created with the ongoing, persistent goal of countering ISIS cyberspace operations.

The term "jointness" is also used by General Nakasone to describe cross-team capabilities or partnerships that enable increased agility in cyberspace. This contrasts with dedicated cyber "service" components, which are embedded within military units like the Army, Navy, Air Force, and Marines, and aid these groups in combat. Vice Admiral Timothy White argues that cyberspace links domains like the air, land, and sea, and has therefore necessitated a "joint" conception (Schneider, 2020). Nakasone further asserts that USCYBERCOM will become increasingly "joint" in future (Joint Force Quarterly Issue 92, 2019).

Notably, the US military has already used "joint" structures against terrorist threats. During the 2007 "Surge" into Iraq, the Special Forces assembled "Intelligence, Surveillance, and Reconnaissance" (ISR) teams with people from a range of agencies – including the FBI, CIA, NSA, and Treasury Department – to target terrorists (Sutherland, 2019). By putting everyone under one roof, these organizations reduced the frictions of cross-agency collaboration. Thus, ISR teams could leverage each agency's strengths on an as-needed basis to do whatever the situation demanded, from tracking fund flows to eavesdropping on phone calls to following a

vehicle. Because of this "jointness" (also called "agile culture"), ISR teams were good at identifying targets for the Special Forces to eliminate. Thus, the success of "joint" teams in kinetic counterterror operations presaged JTF-ARES' success in a cyber offensive.

4. JTF-ARES: A Case Study in Military and Cyber Co-ordination

We now turn to the example of Joint Task Forces-ARES (JTF-ARES), an NSA and U.S. Cyber Command collaboration intended to "deny, degrade, disrupt, and manipulate ISIS' info space" (Temple-Raston, 2019a). The operations of JTF-ARES were inspired by the realization that ISIS materials were distributed through only ten nodes on the internet – by gaining administrative access, the JTF hoped to deny access and disrupt ISIS operations via a mission entitled "Operation Glowing Symphony" (Temple-Raston, 2019a).

After extensive planning and reconnaissance, the offensive portion of "Operation Glowing Symphony" commenced. The task force successfully gained administrative access to ISIS networks and denied access to existing administrators, eliminating a predetermined list of "targets" (Figure 1). Following this initial attack, JTF-ARES completed more subtle follow-on attacks that focused on "eroding morale" (Temple-Raston, 2019a). Examples include deleting ISIS-created content, draining cell phone batteries, and slowing internet for ISIS network users (Temple-Raston, 2019a).

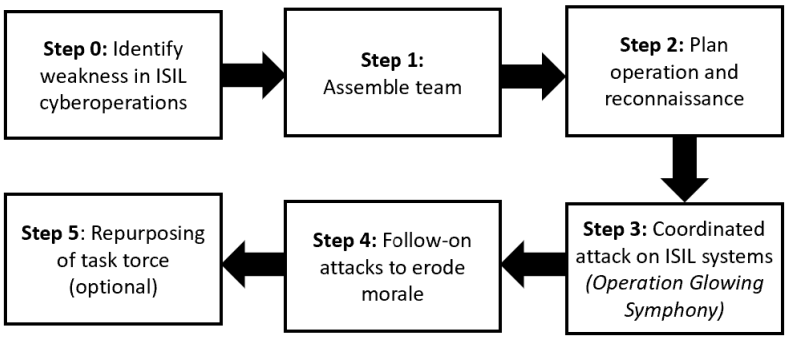


Figure 1: The lifecycle of JTF-ARES, from formation to repurposing

Internal documents deemed the operation a success, as it prompted a delay in official ISIS magazine publications. Other researchers noted a decrease in extremist Twitter activity following the authorization of Operation Glowing Symphony (Martelle, 2018). Alternatively, some have argued that ISIS' continued online presence represents an operational failure (Temple-Raston, 2019a). It is challenging to conclude whether a persistence attack is a distinct "success" or "failure" – persistence operations, by definition, have no end and require continual response.

4.1 JTF-ARES: The First of Its Kind

Before the creation of JTF-ARES, the U.S. Cyber Command launched offensives against ISIS, although no organizational structure for these offensives is available to the public. General Nakasone described cyberoperations against ISIS as "episodic," rather than continuous (Temple-Raston, 2019a). Individual attacks on ISIS servers from USCYBERCOM had no lasting impact on their operations – it is reasonable to conclude that JTF-ARES enabled success in part due to its organizational structure. JTF-ARES's novelty and relative success is also indicated by its longevity: ARES leadership noted that the brand name will be kept despite completion of Operation Glowing Symphony due to the benchmarks set by its approach (Atlantic Council, 2019).

Furthermore, the addition of JTF-ARES appears to be an anomaly within USCYBERCOM. For one thing, it represented the first coordinated effort between the NSA and USCYBERCOM (Temple-Raston, 2019a). It also represented a unique example of a JTF that was independent of a kinetic unit, which stands in contrast to the existing components of the U.S. Cyber Command. Aside from the four service components, the U.S. Cyber Command also includes a Cyber Mission Force (established in 2012), with 133 sub-teams. These teams appear to be primarily defensive (e.g., a Cyber Protection Team), focused on enabling combat (e.g., Combat Mission Force), or response-based (e.g., a National Mission Force to block attacks). Other components like the Cyber National Mission Force (CNMF) represent a larger move toward hybrid offensive and defensive operations, but small and singularly focused task forces are a relative anomaly (Command History, n.d.). The idea that

cyberoperations generally serve kinetic operations can also be seen in the integration of cyberoperations into service groups within the USCYBERCOM, which are created for primarily kinetic purposes. AFCYBER, ARCYBER, and MARFORCYBER are all examples of cyber service subunits that are created to serve their respective branches of military (Air Force, Army, and Marines, respectively).

Most importantly, the task force has been cited by military leadership as establishing new benchmarks in the history of cyberoperations (Atlantic Council, 2019). Brigadier General Leonard Anderson (Deputy Commander of JTF-ARES at USCYBERCOM) noted that singular cyberoperations not in support of a military subunit are a "very rare occurrence," and are more frequently part of a campaign plan that pairs cyberoperations with kinetic operations (Atlantic Council, 2019). General Anderson further notes that current cyber operations generally directly support a marine or soldier on the ground (i.e., kinetic forms of warfare) – while in 15 years, he anticipates cyberoperations that are not singularly in support of a military unit (Atlantic Council, 2019). This perspective of cyberoperations as a tool for accomplishing kinetic missions renders the creation of JTF-ARES unique for two reasons. First, JTF-ARES articulates a unique cyber "end state" in its founding documents, namely that "ISIL is denied use of the cyberspace domain to enable their critical capabilities" (U.S. Cyber Command, 2017). Secondly, this mission statement is not directly in support of a kinetic attack – while weakening ISIS cyberoperations would aid in a long-term kinetic goal, the self-described end state of Operation Glowing Symphony was a cyber, rather than a kinetic goal.

5. Exploring the Counterfactual: Why USCYBERCOM?

It is worth considering briefly why JTF-ARES was created within the purview of the U.S. Cyber Command, as opposed to being conducted within a service component of the USCYBERCOM, or within the Joint Task Force responsible for Operation Inherent Resolve, a coalition of government forces created specifically to counter ISIS (Operations and Exercises, n.d.).

The initial realization that ISIS materials were being distributed through relatively few accounts occurred within the U.S. Cyber Command. It may be the case that had this realization occurred elsewhere, the organizational process and operation would have resulted differently. Had Operation Glowing Symphony been implemented within a military service unit, there may have been conflicts of interest between combat and reconnaissance goals (Stolz, 2021). For example, the cyberoperation may have been apprehended by ISIS, and operations could have been moved to a new network – this may have jeopardized existing military operations that relied on cyber reconnaissance.

Furthermore, Operation Glowing Symphony has been described as the "most complex offensive cyberspace operation" that USCYBERCOM had conducted at the time, and planning documents alluded to the likelihood of conducting similarly complex cyberoperations in future (Martelle, 2020). This complexity necessitated a dedicated team – had the operation been conducted by a service component, or within an existing combat operation, the scale of objectives may have been less ambitious or compromised in favor of other combat-focused goals.

6. Administrative Challenges

JTF-ARES and its Operation Glowing Symphony represent broader moves towards "jointness" and "persistence" doctrines within military conceptions of cyberspace. JTF-ARES inspired similar small and singularly-focused task forces that are able to perform reconnaissance, offensive, and defensive operations. However, JTF-ARES was met with administrative hurdles and organizational challenges that jeopardized the efficacy of Operation Glowing Symphony. Examining these challenges offers insight into future task-force design.

First, Operation Glowing Symphony was met with "non-concurs", or objections from the CIA, State Department, and FBI. Agency leaders felt it was necessary to alert allied countries to potential disruption on servers containing ISIS information (Nakashima, 2017). This complication took weeks to resolve, and prevented the mission from being carried out as designed. Furthermore, extensive clearing processes were required by JTF-ARES to prevent conflicts with ongoing offensive operations in Mosul, Iraq. In FOIA-released documents, the USCYBERCOM notes that (redacted) measures could help "expedite the request and approval process" for clearing targets, especially considering the "likelihood that USCYBERCOM will be conducting more frequent and widely scoped cyber operations…in future" (Martelle, 2020).

Furthermore, the future of JTF-ARES poses questions about whether task forces are easily repurposed, and whether they should be permanent or temporary in nature. JTF-ARES has been refocused to cover great power

competition under the US Indo-Pacific Command's purview, collaborating with MARFORCYBER (Atlantic Council, 2019). The refocusing of JTF-ARES to great power competition as opposed to ISIS cyberoperations suggests that persistence does not mean permanence. A "persistent" task force need not focus on a single threat forever and may pivot in focus.

General Anderson notes that JTF-ARES staffing is dynamic: as JTF-ARES collaborates closer with MARFORCYBER, some of the military personnel within JTF-ARES (which constitute approximately 40% of operators) will return to their own subunits, and the marine-affiliated staffings will increase (Atlantic Council, 2019). Thus, JTF-ARES can be refocused and restaffed according to focus area and may offer increased agility in comparison to service subunits. It also, however, may require specialized training and an increased focus on collaboration to allow civilian and military operators from various agencies to work together.

Staffing and training protocols may need to be customized to enable collaboration between different agencies. FOIA-released documents claim that staffings within JTF-ARES were filled first with NSA employees, then USCYBERCOM service employees (i.e., members of service subunits like MARFORCYBER) (U.S. Cyber Command, 2017). While this protocol offers structure, it also implies that new employees may have different backgrounds and may struggle to collaborate, especially considering JTF-ARES marks the first major collaboration between U.S. Cyber Command and the NSA. The traditional General Staffing System employed by most NATO governments includes separate employees for staffing and training (U.S. Army Special Operations Command History Department, 2011). In the case of JTFs that draw from disparate organizations, staffing and personnel may need to work closely to design custom training courses to ensure personnel can collaborate effectively.

7. Potential Application of Joint Task Forces

JTF-ARES planning documents leave open the possibility of future, similar cyberoperations. We have discussed how the creation of JTF-ARES can be considered something of an anomaly, being the first NSA-USCYBERCOM JTF of its kind. It may, therefore, be useful to consider when a persistent JTF should be created. We propose two key questions to determine whether a JTF may be useful:

Does the cyberoperation require reconnaissance, defensive, and offensive components?

Reconnaissance operations in support of combat missions may be better suited to service subunits of USCYBERCOM. In contrast, JTF-ARES required months of reconnaissance and planning, coupled with offensive and psychological operations to erode morale. Furthermore, JTF-ARES required NSA and U.S. Cyber Command expertise, which meant a JTF was particularly useful.

Does the operation have a cyber (not primarily kinetic) "end state"?

As discussed previously, the concept of a cyberoperation not in service of a kinetic goal is rare. Cyberoperations with self-contained, non-kinetic goals may be ill-suited to service subunits, where their priorities may be subverted. Instead, an individual task force within U.S. Cyber Command that employs "jointness" to recruit specialty talent may be helpful in accomplishing singular, or predominantly cyber goals.

7.1 Further Scenarios and Applications

Having discussed the criteria for establishing a persistent JTF, it is worth considering real-world scenarios that may warrant this organizational structure. As mentioned above, General Nakasone observes that cyberthreats are evolving from episodic, response-based espionage to disruptive, persistent attacks, and this change warrants a new response. Nakasone identifies early cyberthreats as mostly information theft, while modern, "corrosive" threats weaponize information in subtle influence campaigns, a primary example being the 2016 US election (Joint Force Quarterly Issue 92, 2019). A persistent JTF would be especially well-suited to counter this new type of cyberthreat. Indeed, JTF-ARES influenced the formation of the Russia Small Group (RSG), a joint, persistent (and now permanent) task force established by USCYBERCOM and the NSA to counter Russian election influence during the 2018 US midterm elections (Temple-Raston, 2019b; Vavra, 2019).

Election security is particularly well-suited to the JTF structure and it fulfills both of the outlined persistent JTF criteria. To illustrate, one of RSG's public successes was the "hunt forward" initiative, which publicized malware found in foreign government systems. This threat-hunting directly helped foreign governments and had the secondary effect of allowing foreign allies to proactively defend against future malware (Warner, 2020). This mission reportedly allowed for the mass inoculation of "millions of systems" (Warner, 2020). Such activities may be well-suited to a persistent JTF, where a mission like "threat hunting" requires reconnaissance and offensive

capabilities (criterion 1) to investigate compromise on foreign networks. Threat hunting also has little relevance to other kinetic operations, but promotes the long-term health of cyber systems (criterion 2) (Nakasone & Sulmeyer, 2020). Beyond threat hunting, little is known about the operations of the RSG, but General Nakasone also described "offensive cyber and information operations" in connection to the 2018 midterm elections (Senate Hearings, 2015). Thus, the RSG effectively completed reconnaissance and offensive operations under a single, persistent task force structure, and, per Nakasone, "disrupt[ed] those planning to undermine the integrity of the 2018 midterm elections" (Senate Hearings, 2015). Furthermore, a task force dedicated to election security arguably has an end state that transcends kinetic and cyber domains. Senate reports suggest that Russia intended to sow general distrust in democratic voting processes, a goal that included cyber methods but would likely also have caused confusion at physical polling stations (United States Senate Select Committee on Intelligence, 2019). An effective task force would therefore have both cyber and physical goals, quashing DoS and phishing attacks at the local level while simultaneously maintaining the physical order of the election.

The Department of Homeland Security reported that "numerous actors are regularly targeting election infrastructure, likely for different purposes, including to cause disruptive effects, steal sensitive data, and undermine confidence in the election" (United States Senate Select Committee on Intelligence, 2019). Since a single, powerful state has many vectors to attack an election, a dedicated task force may be well-suited to counter states with a motive and history of election interference. While no specific operational successes have been publicized, media outlets deemed the RSG a success in having "deterred some interference during the midterms," and Senator Blumenthal (D-CT) noted (without specifying) that the RSG had "some successes that the American people should know happen[ed]" (Warner, 2022). Such a success suggests that state-specific task forces, established to defend or act in advance of major events (like elections), may be well-suited to a persistent JTF.

7.2 Potential for Future Research

Although the idea of persistence in cyberspace has reached the highest levels of government, there is minimal practical literature on how to deploy persistent Joint Task Forces. The 2019 National Defense Authorization Act ("NDAA") classified "clandestine military…operations in cyberspace" as "traditional military activity." Thus, non-attack operations in cyberspace face fewer reporting requirements to Congress than do covert actions, the latter of which undergoes complex interagency approvals (Chesney, 2018b). The act also preauthorized the Secretary of Defense to undertake "proportional" cyber measures against an "active, systematic, and ongoing campaign of attacks" against the US that originates from Russia, China, North Korea, or Iran (U.S. Congress, 2018). Furthermore, in 2018, President Trump issued National Security Presidential Memorandum-13, which allows cyberoperations to be conducted with DoD approval, as opposed to the previously complex and time-consuming interagency process (Freedberg, 2018; Weingarten, 2020). In sum, the Secretary of Defense has been given increased autonomy and the ability to conduct offensive and defensive operations in accordance with a persistent, "defend forward" strategy (Chesney, 2018a; Chesney, 2019). While this strategic change is evident in legislation, tactical literature has overlooked the administrative opportunities and challenges of persistent Joint Task Forces, as this paper endeavors to do.

The existing literature about joint military operations in cyberspace does not adequately explain how to deploy tactical cyber operations on a day-to-day operational level. For example, the US Army War College's Strategic Cyberspace Operations Guide, a key document for students hoping to understand cyberoperations' structure, offers high-level information on conducting cyberoperations and potential "big picture" challenges, but fails to offer details on deploying cyberoperations day-to-day (US Army War College, 2022). The document references "defend[ing] forward" as a key and novel approach to cyberspace operations, but it does not discuss how persistent JTFs may differ in challenge, scope, or operation from existing organizational structures. Similarly, the 2018 U.S. Cyber Command's high-level "vision" document articulates persistent engagement as a key priority, citing lengthy approval processes as a hindrance that can force the US into a disadvantageous, "reactive" mode (U.S. Cyber Command Vision, 2018). Future research could discuss the potential of combining several discrete functions into a single JTF to reduce communication overhead costs, or how to effectively repurpose a persistent JTF upon mission completion.

8. Conclusion

This paper discussed the potential for persistent Joint Task Forces that address singularly cyber-oriented goals, using the case study of JTF-ARES. A history of the U.S. Cyber Command was presented, and early conceptions of

cyberoperations as "response"-based have been explained. Further, the terms "joint" and "persistent" have been defined as emerging cyberspace doctrines within the US Government. The unique creation of JTF-ARES has been explored, and we suggest that Operation Glowing Symphony was more likely to have failed under a different military-oriented or cyber service unit. Thus, the usefulness of Joint Task Forces has been shown, and criteria for their suitability has been established for future operations. We also discussed hypothetical scenarios where a persistent JTF may be useful, including election security, especially against a state with a history of cyber interference. While this paper has offered some criteria that may be useful in determining whether a JTF is necessary, further examinations of organizations like the Russia Small Group may offer interesting case studies. The doctrines of persistence and jointness are relatively new compared to response-oriented warfare concepts and merit further research, especially in the day-to-day practicalities of staffing and deployment.

References

Atlantic Council (2019) *Cyber Operations in Context: A Look at Joint Task Force Ares*, YouTube. YouTube. Available at: www.youtube.com/watch?v=aQsDTvQf4_E (Accessed: 9 February 2023).

Chesney, R. (2018a) *The 2018 DOD Cyber Strategy: Understanding 'Defense Forward' in Light of the NDAA and PPD-20 Changes*, Lawfare Blog. Available at: https://www.lawfareblog.com/2018-dod-cyber-strategy-understanding-defense-forward-light-ndaa-and-ppd-20-changes (Accessed: April 17, 2023).

Chesney, R. (2018b) *The Law of Military Cyber Operations and the New NDAA*, Lawfare Blog. Available at: https://www.lawfareblog.com/law-military-cyber-operations-and-new-ndaa (Accessed: 17 April 2023).

Chesney, R. (2019) *Covert Military Information Operations and the New NDAA: The Law of the Gray Zone Evolves*, Lawfare Blog. Available at: https://www.lawfareblog.com/covert-military-information-operations-and-new-ndaa-law-gray-zone-evolves (Accessed: April 17, 2023).

Command History (n.d.*) U.S. Cyber Command.* Available at: www.cybercom.mil/About/History/ (Accessed: 9 February 2023).

Cyber Mission Force (2022) *U.S. Army Cyber Command.* Available at: www.arcyber.army.mil/Resources/Fact-Sheets/Article/2079661/cyber-mission-force/ (Accessed: 9 February 2023).

Cyber National Mission Force Public Affairs (2022) The Evolution of Cyber: Newest Subordinate Unified Command is nation's Joint Cyber Force, U.S. Cyber Command. Cyber National Mission Force Public Affairs. Available at: https://www.cybercom.mil/Media/News/Article/3250075/the-evolution-of-cyber-newest-subordinate-unified-command-is-nations-joint-cybe/ (Accessed: 16 April 2023).

Freedberg, S.J. (2018) *Trump Eases Cyber Ops, But Safeguards Remain: Joint Staff*, *Breaking Defense*. Available at: https://breakingdefense.com/2018/09/trump-eases-cyber-ops-but-safeguards-remain-joint-staff/ (Accessed: April 17, 2023).

Inamete, U.B. (2022). *The Unified Combatant Command System*. Expeditions with MCUP digital journal. Marine Corps University Press. https://www.usmcu.edu/Outreach/Marine-Corps-University-Press/Expeditions-with-MCUP-digital-journal/The-Unified-Combatant-Command-System/ (Accessed: 16 April 2023).

Integration and Synchronization of Joint Fires (July 2018). Retrieved November 27, 2022, from www.jcs.mil/Portals/36/Documents/Doctrine/fp/int_and_sync_jointfires.pdf?ver=2018-%2009-18-102801-350.

Joint Force Quarterly Issue 92 (2019) National Defence University Press (NDU Press). Available at: ndupress.ndu.edu/JFQ/Joint-Force-Quarterly-92.aspx (Accessed: 9 February 2023).

Martelle, M. (ed.) (2020) USCYBERCOM *After Action Assessments of Operation Glowing Symphony,* National Security Archive. Available at: nsarchive.gwu.edu/briefing-book/cyber-vault/2020-01-21/uscybercom-after-action-assessments-operation-glowing-symphony (Accessed: 9 February 2023).

Martelle, M. (ed.) (2018) *Joint Task Force Ares and Operation Glowing Symphony: Cyber Command's Internet War against ISIL,* National Security Archive. Available at: nsarchive.gwu.edu/briefing-book/cyber-vault/2018-08-13/joint-task-force-ares-operation-glowing-symphony-cyber-commands-internet-war-against-isil (Accessed: 9 February 2023).

Nakashima, E. (2017) *U.S. Military Cyber Operation to Attack Isis Last Year Sparked Heated Debate Over Alerting Allies,* The Washington Post. WP Company. Available at: www.washingtonpost.com/world/national-security/us-military-cyber-operation-to-attack-isis-last-year-sparked-heated-debate-over-alerting-allies/2017/05/08/93a120a2-30d5-11e7-9dec-764dc781686f_story.html (Accessed: 9 February 2023).

Nakasone, P.M. and Sulmeyer, M. (2020) *How to Compete in Cyberspace*, *Foreign Affairs*. Available at: https://www.foreignaffairs.com/articles/united-states/2020-08-25/cybersecurity?check_logged_in=1&utm_medium=promo_email&utm_source=lo_flows&utm_campaign=registered_user_welcome&utm_term=email_1&utm_content=20230416 (Accessed: April 16, 2023).

Operations and Exercises. (n.d.). *U.S. Cyber Command*. Retrieved November 27, 2022, from www.centcom.mil/OPERATIONS-AND-EXERCISES/OPERATION-INHERENT-%20RESOLVE/

Proctor, A. (2006) *AF launches Cyberspace Task Force, Air Force.* Air Force. Available at: www.af.mil/News/Article-Display/Article/131398/af-launches-cyberspace-task-force/ (Accessed: February 9, 2023).

Schneider, J.G. et al. (2020) *Ten years in: Implementing strategic approaches to cyberspace, U.S. Naval War College Digital Commons.* Available at: digital-commons.usnwc.edu/usnwc-newport-papers/45/ (Accessed: February 9, 2023).

2015 Senate Hearings*, Department of Defense Authorization for Appropriations for Fiscal Year 2016 and the Future Years Defense Program: Hearing on S. 1376 before the S. Comm. on Armed Servs., Subcomm. on Intelligence, Emerging Threats and Capabilities*, 114th Cong. 415 (2015) (statement of Admiral Michael S. Rogers, Commander, US Cyber Command)

United States Senate Select Committee on Intelligence. (2019). *Report of the Select Committee on Intelligence United States Senate on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election*. Retrieved from https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume1.pdf (Accessed: 16 April 2023)

Stolz, M. (2021). *Competing Interests of Cyberintelligence and Cyberdefence Activities in Neutral Countries*. 9.

Sutherland, J.V. (2019) in Scrum: *The Art of Doing Twice the Work in Half the Time*. Random House, pp. 54–58.

Temple-Raston, D. (2019a) *How the U.S. Hacked Isis, NPR.* NPR. Available at: www.npr.org/2019/09/26/763545811/how-the-u-s-hacked-isis (Accessed: 9 February 2023).

Temple-Raston, D. (2019b) *Task Force takes on Russian Election Interference*, *NPR*. NPR. Available at: https://www.npr.org/2019/08/14/751048230/new-nsa-task-force-takes-on-russian-election-interference (Accessed: 16 April 2023).

U.S. Army War College. (2022). *Strategic Cyberspace Operations Guide.* [online] Available at: https://csl.armywarcollege.edu/USACSL/Publications/Strategic_Cyberspace_Operations_Guide.pdf (Accessed: 16 April 2023).

U.S. Army Special Operations Command (USASOC) History Department (2011) *The General Staff System, U.S. Army Special Operations Command History Office.* Available at: www.arsof-history.org/articles/v7n2_general_staff_system_page_1.html (Accessed: 9 February 2023).

U.S. Congress. (2018*). John S. McCain National Defense Authorization Act for Fiscal Year 2019.* [online] Available at: https://www.congress.gov/115/plaws/publ232/PLAW-115publ232.pdf (Accessed: 16 April 2023).

U.S. Cyber Command. (2017*). Joint Task Force Areas. [pdf]* U.S. Strategic Command. Available at: https://www.stratcom.mil/Portals/8/Documents/FOIA/FOIA%2017-023,%2017-033,%2017-064%20-%20USCYBERCOM%20Joint%20Task%20Force%20Areas.pdf?ver=2017-04-19-111941-797 (Accessed: 17 April 2023).

U.S. Cyber Command. (2018). U.S. Cyber Command Vision April 2018. [online] Available at: https://www.cybercom.mil/Portals/56/Documents/USCYBERCOM%20Vision%20April%202018.pdf (Accessed: 16 April 2023).

Vavra, S. (2019) NSA's Russian Cyberthreat Task Force is Now Permanent, CyberScoop. Available at: https://cyberscoop.com/nsa-russia-small-group-cyber-command/ (Accessed: 16 April 2023).

Warner, M. (2022). *U.S. Cyber Command's First Decade.* In M. Warner, The United States' Defend Forward Cyber Strategy (pp. 33–64). Oxford University Press. doi.org/10.1093/oso/9780197601792.003.0004

Weingarten, D. (2020) *Congress receives long-awaited memorandum from White House on Cyber Policy , MeriTalk*. Available at: https://www.meritalk.com/articles/congress-receives-long-awaited-memorandum-from-white-house-on-cyber-policy/#:~:text=National%20Security%20Presidential%20Memorandum%2013,offensive%20cyber%20operations%20are%20approved. (Accessed: 16 April 2023).

Cyber Power in the African Context: An Exploratory Analysis and Proposition

Petrus Duvenage, Wilhelm Bernhardt and Sebastian von Solms
University of Johannesburg, South Africa

duvenage@live.co.za
wbernhardt@vodamail.co.za
basievs@uj.ac.za

Abstract: While the centrality of cyber power in the safeguarding and advancing nation states' national interests and objectives is now widely accepted, the academic discourse (on cyber power) is still incipient. In literature reviewed, cyber power is predominantly viewed as comprising of two dimensions, namely offensive and defensive. The exploratory analysis we conducted found that Africa's unique, contextual factors necessitate an expanded conceptualisation of cyber power. This alternative conceptualisation does not dispute the existing notion that cyber power has offensive and defensive dimensions. The fact that cyber is by its very nature borderless and that African countries function in an interconnected global arena of competition and conflict, are also not contested. What is required is the addition of a third dimension to cyber power, namely developmental power. This paper advances a tentative proposition on a cyber-power triad (with offensive, defensive and developmental dimensions). This proposition, we argue, is more apposite to African countries' national objectives —strategically and in the allocation of resources. At least on a notional level, the cyber-power triad can guide the leveraging of the asymmetric advantages that cyber space offers African nation states and in a manner that pursues all three (cyber power) dimensions in a complementary manner. Such synergetic wielding of cyber power is one of the keys indispensable to African countries addressing their substantial challenges and unlocking their vast potential.

Keywords: Cyberpower, Nation States, Developing Countries, Cyber Power Model, Cyber Power Index, Global South.

1. Introduction

It is projected that Africa's population will double as early as 2050, characterised by a sizeable youth bulge. In fact, according to the World Economic Forum (WEF), by 2030, young Africans are expected to constitute 42% of global youth (El Habti, 2022). The political and economic implications of this trend are multiple, and of considerable impact. On the one hand, it raises fears of burgeoning unemployment, rising poverty and pervasive conflict, whilst on the other there is tremendous potential for economic development and mounting political influence, not only on the continent, but on a global scale. The African Union's (AU) Digital Transformation Strategy for Africa (2020 – 2030) acknowledges these projections, and notes that 'digital transformation' is a driving force for sustainable growth (AU, 2020). It envisages an integrated and inclusive digital society and economy in Africa that improves the quality of life of Africa's citizens, strengthen the existing economic sector, enable its diversification and development, and ensure continental ownership with Africa as a producer and not only a consumer in the global economy (AU, 2020). It is therefore not surprising that the optimal utilisation of cyberspace features prominently as an enabling force in the AU's 'Agenda 2063', promoted as Africa's "blueprint ...for transforming Africa into the global powerhouse of the future" (AU, 2015). Optimising the utilisation of cyber space, in turn, and as this paper will show, has as prerequisite a fit-for-purpose conceptualisation of cyber power.

The academic discourse on cyber power is in various respects still incipient and evolving (Cavelty, 2018; BAE, 2022). At least in as far as consultant literature is concerned, 'cyber power' emerged as a distinctive concept in the United States (US) early in the first decade of the 21st century, but only gained wider academic traction from 2009 onward (Starr, 2009; Cavelety, 2018). The limited academic literature that has subsequently been advanced is overwhelmingly predisposed directed to the Global North. (The 'Global North', when used in this paper denotes the Western world and other developed economies, including Japan and Israel. The Global South largely corresponds with developing countries in Africa, Latin America and the East (Pagel *et al,* 2014).

While there are differences, the limited literature on cyber power mostly reflects Global North commonalities in approach and assumptions. Unsurprisingly, this literature is predisposed towards the national objectives of the Global North's powerful states and that of their major adversaries. Accordingly, cyber power was, and for a

substantial part is still, analysed in the context of inter-state conflict, specifically in relation to economic and especially military type antagonism and rivalry. As recently as 2018, the Global North's conceptualisation of cyber power predominantly remained one dimensional, namely 'offensive' (Cavelty, 2018). Of late, however, there is growing acceptance that 'cyber power' has an integral second dimension, namely defensive power. In the Global North's view this sum of an actor's offensive and defensive cyber power is relative to those of others (Devanny, Goldonia and Medeiros, 2022). There is thus a hierarchal distinction to be made between actors with more, lesser or equal cyber power.

Even in comparison with the limited (but expanding) Global North literature, peer-reviewed academic contributions which specifically explores cyber power within the context of sovereign states located towards the lower end of the international spectrum of national power, is scarce (van Niekerk, 2019). More often than not, cyber power concepts and indices from the Global North are applied practically without qualification to the Global South in general, and Africa in particular.

Herein lies the glitch - as with other forms of power, conceptualisations of cyber power are context specific and cannot necessarily be applied unchanged (if at all) to other polities (Cavelty, 2018). Subsequently, this paper poses the following questions:

- What, in more detail than outlined above, is the Global North's mainstream view of cyber power?
- How pertinent and applicable is the global North's conceptualisation of cyber power to the global South in general, and Africa in particular?
- Is there a need for an expanded or alternative conceptualisation of cyber power?
- If indeed required, what contextual factors should be considered should in shaping an African specific conceptualisation?
- Albeit tentative, what proposition on cyber power apposite can be advanced?
- What are the implications of an alternative proposition for the assessment of cyber power in Africa?

This paper's central contention is that Africa's unique, contextual factors necessitates an alternative conceptualisation of cyber power, which should be more inclusive, holistic, and incorporate not only the ability to defend the state's interests, and/or to influence or attack an adversary with cyber instruments, but critically also a third dimension: the capacity to optimise the state's development policies. Thus, we advance a tentative proposition on a cyber-power triad (with offensive, defensive and developmental dimensions). This proposition, we argue, is more relevant to African countries' strategic national objectives. At least on a notional level, the cyber-power triad can guide the leveraging of the asymmetric advantages that cyber space offer African nation states and in a manner that pursues all three (cyber power) dimensions in a complementary manner. In this way the relatively limited resources and capacity of African countries can be optimised.

This paper is categorically qualified as an exploratory examination and tentative proposition. It is the first of a series of papers flowing from research at the University of Johannesburg (UJ) on the topic of cyber power in the Africa context. This being the first paper, we endeavour to lay a notional foundation that is being concretised in follow-up research.

Structurally, the paper is aligned with the research questions and consists of the following:

- Section 2 provides an overview of the predominating conceptualisation of cyber power in the Global North.
- Section 3 examines the imperative of developing an expanded proposition on cyber power.
- Section 4 identifies the factors to be considered in formulating a proposition on African cyber power.
- Section 5 advances an African cyber-power triad as a tentative proposal.
- Section 6 examines the implications of this proposal on assessing cyber power in Africa.
- Section 7 concludes the paper by outlining ongoing research on this topic.

2. The Global North's conceptualisation of cyber power

Within the context of international relations, cyber power can broadly be defined as that element of national power aimed at protecting and advancing national interests through and in cyberspace (Devanny 2021; Voo, Hemani and Cassidy 2022). Because this definition is so broad, it is, for all intents and purposes 'compatible' with all works consulted for purposes of this paper.

However, it is on moving from this broad definition to more specific and granulated conceptualisations of cyber power, that views diverge. It is beyond the scope of this paper to dissect these variations. The aim here is to outline the Global North's dominant view of cyber power in the academic discourse. This view is, for a large part, articulated in literature from Western countries. Although it admittedly risks oversimplification, for our purposes here, the Global North essentially views cyber power as comprising of two dimensions, namely the 'offensive' (wielding of power) and the defensive (cybersecurity and resilience). The said two dimensions are further explained in subsections 2.1 and 2.2. This is followed by a concise overview of current cyber power indices, predicated on these dimensions (Subsection 2.3).

2.1 'Offensive' dimension

This Global North's dominant view posits cyber power as an integral part of a nation state's power contestation with other states and non-state actors (Nye 2010). Cyber power is seen both as a dimension of state power and the 5th domain of warfare (in addition to land, sea, air and space). Cyber power is furthermore asserted as cutting across, all *four elements* of national power (Political/Diplomatic, Information, Military and Economic) (Kuehl 2009). As national power in general, cyber power is key to gaining an advantage over competing actors and influencing the environment (Sheldon, 2011). This is achieved through a combined application of coercion ('hard power') and persuasion ('soft power'). Soft and hard power are not exclusive 'categories' but rather comprises a continuum of overlapping aspects ranging from diplomacy and intelligence gathering, to cyber weapons (Van der Waag-Cowling 2019; Devanny, Goldoni & Medeiros, 2022). Used in unison, soft and hard power constitute smart power'. We note that, in a narrow definitional sense, the 'wielding' if cyber power is not always of an 'offensive' nature. 'Offensive' as used here involves more than cyber war /disruptions and also includes influencing, intelligence gathering and surveillance. In fact, "mature offensive cyber capability" includes "intelligence collection and analysis" (Devanny, Goldoni and Medeiros, 2022), as well as cyber counterintelligence (CCI). The latter in turn includes substantial defensive facets (Duvenage, 2019).

2.2 'Defensive' dimension

Cavelty (2018) rightly points out that the Global North's discourse was until very recently all about the offensive domination of one actor over others– be it through hard and/or soft power. Cyber power was thus up to 2018 mainly a one-dimensional construct. One notable exception has been Klimburg's model (2011), which:

> "[Takes] cyberpower out of the context of projecting national interests coercively and adds a strong defensive cyber-security element, built on a best-practice model of cyber-security policy. Such a view highlights that cyberpower and cyber-security are intricately linked. ***A political entity possesses cyber-power if it has the ability to shape aspects of the global cyber-security landscape***. However, a cyber-power also needs to be able to 'defend' against cyber-threats, or rather, manage them adequately. These necessities, internal cyber-resilience and external cyber-power, build on each other: in this view, there cannot be any true cyber-power without cyber-resilience – and vice versa." (Cavelty, 2018 - emphasis added)

Currently, cybersecurity is no longer viewed as complimentary to, but as an integral second dimension of, cyber power. Also this dimension emphasises the importance of projecting power globally. This two-dimensional view of cyber power is clearly reflected in Global-North instruments that aim at measuring cyber power. These instruments are discussed in the next subsection.

2.3 Relative nature of cyber power

As was noted earlier, the Global North posits cyber power as hierarchal and relative. Respective nation states thus possess more, less or roughly equal degrees of cyber power. Since cyber power is relativistic and relational (Devanny, Goldoni and Medeiros, 2022) distinctions are made between cyber super powers, major powers and lesser role-players (Segal, 2016).

Several instruments (models and indices) exist for assessing cybersecurity. Some of these, such as the Global Cybersecurity Index of the International Telecommunication Union (ITU, 2021a) and the University of Oxford (2021) Cyber Security Capacity Centre's Cybersecurity Capacity Maturity Model for Nations (2021) are well-established, have been applied for several years and have wide global coverage. Literature reviewed, however, identified only two current instruments that assesses cyber power (in its offensive and defensive dimensions), namely the Harvard's Belfer Center for Science and International Affairs index (Voo, Hemani and Cassidy, 2022)

and the United Kingdom-based International Institute for Strategic Studies assessment model (IISSS 2021). Read in conjunction with the BAE's (2022)"hallmarks" of a leading cyber power, these instruments clearly reflects the key contentions of the Global North's view on cyber with power: in order to be a major player a nation state has to project cyber power with a global reach – offensively and defensively. In tabulated format, the measuring criteria employed are as follows:

Table 1: Comparative assessment of cyber power

	Belfer Center	International Institute for Strategic Studies	BAE
Metho-dology	• Quantitative and qualitative	• Quantitative	• Qualitative
Aspects	• Surveilling and Monitoring Domestic Groups • Strengthening and Enhancing National Cyber Defenses • Controlling and Manipulating the Information Environment • Foreign Intelligence Collection for National Security • Growing National Cyber and Commercial Technology Competence • Destroying or Disabling an Adversary's Infrastructure and Capabilities • Defining International Cyber Norms and Technical Standards • Amassing Wealth and/or Extracting Cryptocurrency	• Strategy and doctrine • Governance, command and control • Core cyber-intelligence capability • Cyber empowerment and dependence • Cyber security and resilience • Global leadership in cyberspace affairs • Offensive cyber capability	• Power projection through showing international leadership in national cyber defence capabilities, supported by the cyber defence ecosystem. • A clear and holistic cyber strategy for the nation, led by a strong national cyber agency co-opting all parts of government, industry and society. • A strong technology sector that drives prosperity, enables and contributes to national cyber defences, and is capable of exporting cybersecurity capabilities. • Effective collaboration between government and the technology sector, to influence the evolution of tehnology and standards. • A clearly articulated vision for the future of the internet that embodies the values of government and society. • Cyber diplomacy and relationship-building – the ability to show leadership and promote its vision for norms and values in and through cyberspace. • Responsible development of and use of offensive cyber capabilities.

Sources: Voo, Hemani and Cassidy, 2022; IISS, 2021; BAE, 2022 [Verbatim].

The afore-mentioned instruments are far more selective than the ITU (2021a) and Oxford (2021) cyber security indices with the IISS (2021) assessing 15 and the Belfer Center 30 nation states respectively (Voo, Hemani and Cassidy (2022). Overall, these cyber power assessment instruments find the world's major and superpowers as aspiring to an exceptional degree of offensive and defensive cyber power to achieve their global ambitions and influence. Therefore, such nation states are regularly accorded high rankings.

No African nation state features in the IISS (2021) assessment, while the Belfer Center ranks Egypt as 24th out of 30 countries (Voo, Hemani and Cassidy, 2022). While both these instruments admit limitations, both are commendable in appraising cyber power; ***if*** (i) cyber power is deemed as only having two dimensions – offensive and defensive; and (ii) a nation state aspires to substantially extend its global reach. As will be elaborated upon in the next section, not all nation states have such aspirations and are more focused on pursuing strategies in which 'internal/domestic' national interests dominate. This raises the question of the strategic usefulness of the two-dimensional cyber power construct in the broader setting of less well-resourced nation state's national power, national security and national interests.

3. The imperative of developing an expanded conceptualisation of cyber power

A recent comprehensive study conducted by Çifci (2022) developed a conceptual framework for the comparison of 11 cyber power and cyber security indices. Çifci (2022: 10) rightly asserts that "cybersecurity and cyber power indices, do not necessarily have to be the same or universally accepted. Every individual or organization has their own views on these concepts." The context-specific qualification does of course not distract from the value of

current national-level cybersecurity capability assessments in informing the bolstering of defensive cyber power. Neither does it dispute the insights offered by current assessments of offensive cyber power.

However, since cyber power is contex-specific it needs to be asked whether outside the Global North (i) the construct of cyber power as having two dimensions (offensive and defensive) is sufficient; and (ii) to what degree global aspirations should universally be a major criterion for assessing cyber power? Is it analytically credible to, for a significant part, measure cyber power as states' ability to establish control and exert influence over adversaries within and through cyber space, as though the control and influence are objectives in themselves? Furthermore, whilst it may make strategic sense for some Global North states to develop a vast national capacity of which the sole aims are cyber disruption, cyber espionage, and cyber degradations as the means to influence a rival actor (Valeriano *et al* 2018), for others, with different or less adversarial national objectives, this would actually amount to a strategic misalignment of resources.

The vast majority of African states - the world's 'minor powers' from a national power point of view - generally do not reflect similar levels of global aspirations. The national objectives of such states, and their applications of national power, in fact tend to be located much closer to home, at the intra-state level, and directed towards efforts aimed at eradicating or alleviating poverty by means of development programmes (Odeh, 2010). Development, in this sense, essentially refers to the process through which a country increases its capacity to meet its citizens' basic human needs and raise their standard of living (Kegley and Wittkopf, 1999). Does this emphasis on development imply that cyber power is of lesser strategic importance to African states? Does cyber power mean the same to these states as it does to the Global North?

It is the contention of this paper that African states have unique national objectives that are not a mirror image of those of the world's super and major powers. However, this certainly does not imply Africa is irrelevant from a cyber-power perspective, and only become relevant from the perspective of cyber power indices when and if they pursue cyber power on the same terms as the world's major and superpowers. It is consequently proposed that questions regarding the relative cyber power of the world's nations should be also approached from a complementary perspective, *viz*.: *whether the extent of cyber power at any state's disposal is properly aligned to and sufficient to address that particular state's unique combination of national objectives?* To achieve this, an expanded proposition of cyber power which considers the African context, is required.

4. Context for the conceptualisation of an African cyber power construct

Elaborating on Section 3, this section discusses in more detail the contextual factors that will shape an expanded conceptualisation of cyber power apposite to Africa.

4.1 Political and socio-economical context

States that identify with the Global South are located outside Europe and North America, and mostly tend to be low- to middle-income countries, who are in the process of industrialisation, and are frequently former subjects of colonialism (Pagel *et al,* 2014). All nation states on the African continent are thus categorised as part of the Global South. As the Global South in general, African countries is likely to be disadvantaged in two distinct ways: firstly, in terms of their relative national power deficits *vis-à-vis* the world's more advanced and powerful states; and secondly with respect to their comparative lower levels of socio-economic and political advancement.

From a **national power perspective** African, the respective nation states cannot achieve security in the international arena by relying only on their own capabilities alone, and prefer policies of neutrality or alliances; do not have the necessary military power to project power on a global scale; and favour a high degree of participation in and support for international organisations, where they pursue 'moral' or 'normative' policy positions, and encourage formal rules in order to curb the great powers and strengthen their own position. These states often display risk averse behaviour in international politics, due to the risk of punitive reprisal should they challenge more powerful states; whilst their geographic demands are predominantly restricted to their own and immediately adjacent areas (Toje, 2010). Their overall behaviour in their interaction with other states are in the main characterised by a general and continuing reluctance to coerce and a tendency to promote multilateral, non-military solutions to security challenges (*cf* Fox, 1959). African states are more likely to be embroiled in military or diplomatic concerns about, or conflict with, their immediate neighbours, instability in the region, and troubles somewhere on the continent. Africa also contains stats that are considered fragile, failing or failed, implying that the continuity of the state can, be at risk from internal, domestic threats such as

civil war, revolutions, or insurrections. These countries are generally also far more likely to experience high frequency social instability, violent protests and unrest as well deep-seated ethnic or racial divisions. Overall, they are likely to be significantly vulnerable to erosion of the state's monopoly on violence – a risk that is less unlikely to occur in the major powers. The establishment of a domestic security and intelligence capacity to counter such phenomena, or to collect information about them locally and abroad – also in cyberspace – should therefore make as much strategic sense to the world's minor powers, as it does for their stronger counterparts. Cyberspace has indeed in this respect: *"ma[d]e it easier to subvert and harder to govern*" (Betz and Stevens, 2011).

From a **socio-economic and political development perspective** African states more often than not tend to experience significant developmental challenges that are not normally associated with more powerful, advanced states. As pointed out by Dados and Connel (2012), the 'Global South', of which Africa is an integral part, is more than a metaphor for underdevelopment, as it actually references: "...an entire history of colonialism, neo-imperialism, and differential economic and social change through which large inequalities in living standards, life expectancy, and access to resources are maintained". In the main, the national objectives of African states are not dominated by globally oriented security concerns. Instead, a sizeable part of such states' national power is directed inward, towards intra-state objectives, such as the development and prosperity of the state; the expansion of its economy; the delivery of basic services to all citizens; anti-crime and corruption measures, as well as the institutionalisation and projection of its national values, nation-building and social cohesion.

Probably the most outstanding characteristic of African nation states in the global context is their national power deficit. As will be substantiated in Sections 5 and 6, cyber space offers asymmetric advantages to more effectively address political and socio-economic priorities and concurrently narrow national power deficits. A proposition on leveraging of cyber space's asymmetric advantages also needs to consider African cyber capacity, challenges and constraints. These aspects are overviewed briefly in the next section.

4.2 Cyber context: Potential, capacity, challenges and constraints

African cyber capacity, challenges and constraints are addressed rather comprehensively in subsequent research that follows this paper (see Section 7). Therefore, we suffice here with the following concise, bulleted overview:

- Although there is a wide variance between countries in respect of cyber capacity and maturity, Africa's vast cyber potential is unfolding (AU 2015; AU 2020). The continent has more than 500 million Internet users – putting it ahead of regions such as North America, South America, and the Middle East (Interpol 2021). While it is underserved in terms of Internet coverage capacity and bandwidth speed, Africa has the fastest-growing telephone and Internet networks globally (Interpol 2021).
- The vast majority of African countries are still in early the early stages of transforming to digital economies (AU 2015; AU 2020: ITU 2021b). Concurrently, there is low self-sufficiency in technological innovation and development (Van der Waag-Cowling, 2019).
- From a defensive perspective, most African countries' cybersecurity and resilience are alarming weak in comparison to other continents. The ITU (2021a) for one, assesses African counties overall as the least committed to cybersecurity. Moreover, leadership in respect of collective defence and security, by bodies such as the AU, is weak (Van der Waag-Cowling, 2019).
- Africa is highly targeted by cybercriminals who perceives it to be the "soft underbelly" of global business networks (McCarthy, 2022). Simultaneously, Africa does not escape targeting by state-sponsored Advanced Persistent Threat (APT) actors (Allen and van der Waag-Cowling, 2021).
- The immature offensive capacity that does exist, relies on imported tools and the outsourcing of functions/services. Overall cyber warfare is not "viewed as priority" (Van der Waag-Cowling, 2019). However, several African countries are known to have vendor-sourced cyber espionage and surveillance capacity that can be deployed domestically and externally (Allen and La Lime, 2021).
- Even if measured against very modest foreign policy aims, African countries' cyber diplomacy effort are inadequate (Van der Waag-Cowling, 2019).

African counties are thus faced with a combination of, on the one hand, political and social-economic challenges and, on the other hand, significant cyber constraints. For cyber power to make strategic sense for such states, it must enable them to achieve **asymmetrical advantages** in the context of their national objectives. Asymmetry, in its use here, refers to the ability of a state to enhance or augment any component of its national power through the strategic application of cyber power, in order to gain specific advantages. Practically speaking, it is

thus about the use of cyber power in a manner that achieves **a force multiplication, or the resource optimisation effect**. A tentative postulation on a cyber-power triad providing a theoretical premise for such power optimisation is advanced in the next section.

5. Tentative postulation on African cyber power triad

Our tentative postulation on African cyber power that would enable optimising resources in levering asymmetric advantages, is depicted in Figure 1. This proposition, which is unpinned by the African-specific context provided in Section 4, is holistic in the sense that it encapsulates three forms of cyber power, namely:

- **Dimension 1**: **Offensive cyber power** that (in its loose connotation) refers not only to the capacity to wage cyber war and disrupt, but also the state's ability to utilise cyber means for influencing, intelligence gathering and surveillance. Although 'cyber war is not a priority to most on the continent, the national security of all African countries demands capacity for surveillance and intelligence gathering.
- **Dimension 2: Defensive cyber power,** since the ability to safeguard cyber-related national assets ought to be an indispensable and integral component of cyber power of all African countries – regardless of economic size and the level of digital development.
- **Dimension 3: Developmental cyber power**, which denotes the capacity to pursue national developmental objectives in cyberspace (in a manner that is in tandem and synergy with the defensive and offensive dimensions).

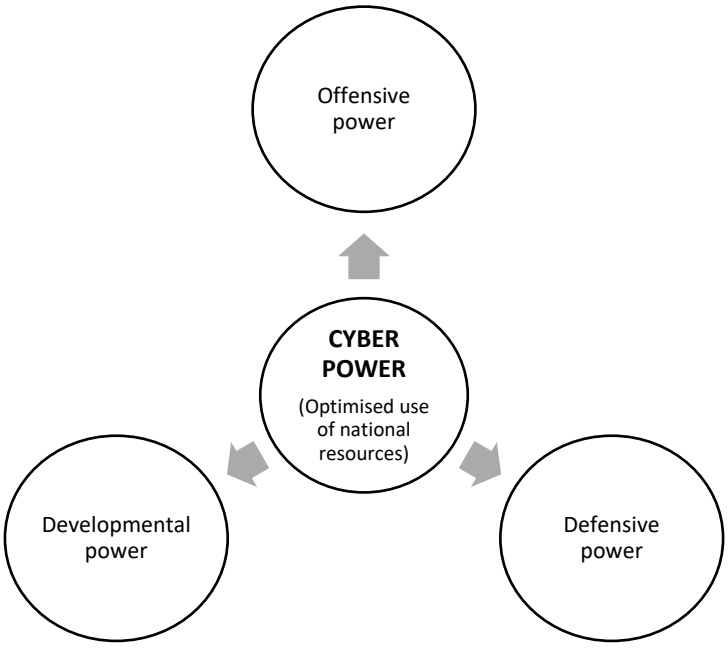


Figure 1: A cyber-power triad for application in the African context

In contrast to cyber power's offensive and defensive dimensions, the paper showed the developmental power dimension as practically unexplored in academic literature (see Sections 2 and 3). As was noted in the introduction, this paper is of an exploratory nature and advances only a tentative proposition. Further work aimed at a substantiated proposal on the developmental cyber power dimension is earmarked for subsequent research (see Section 7). Nevertheless, the following **tentative and preliminary postulations on developmental cyber power** are required to support this paper's contention on the conceptual credibility of a cyber-power triad:

- Developmental cyber power, in its broader connotation, is about a country's collective cyber capacity that is relevant, or potentially relevant, to its national developmental priorities. This capacity has human, institutional, technical and infrastructural facets.

- Like the offensive and defensive dimensions, developmental cyber power depends on appropriate national strategy objectives and measures as well as the allocation of national resources and effort. Given limitations in national resources, African countries have to optimise scarce resources. Therefore, resources can not be devoted to developmental power exclusively. Optimising cyber power requires resources to be shared, and as far as possible serve all three cyber-power dimensions.
- Defensive and offensive cyber power dimensions are essential to support the developmental dimension and *vice versa*.

This section provided a high-level outline of our exploratory postulation on a cyber-power triad for Africa. The next section examines the tentative implications of utilising the cyber-power triad as the core of a qualitative (cyber power) assessment framework.

6. Tentative implications of the cyber-power triad as the core of qualitative assessment framework

The qualitative assessment framework to be designed around the cyber-power triad is predicated on our contentions on national power in Sections 3 and 4. To recapitulate: the national power of states are not equal, their national objectives differ, and therefore their cyber power requirements and strategies should be unique. What works for a superpower as far as national and cyber power are concerned, will most likely not be a usable blueprint for an African country. Therefore, our envisaged approach to assess cyber power is holistic in the sense that it encapsulates all three forms of cyber power and can be applied to evaluate any state in Africa on an equal basis.

Instead of measuring and ranking states in accordance with a presumed hierarchy of greater and lesser cyber powers, the purpose of the holistic assessment would be to ascertain whether any given nation's cyber power is optimally configured to reflect that individual state's cyber power requirements – in other words, whether the state has the ability to do something strategically useful in cyberspace (Gray, 2013). The result of the assessment will be a cyber-power profile, rather than a hierarchical, comparative 'ranking' of states. In this manner, states that require only limited offensive cyber power, are not 'penalised' or relegated to perpetual low positions on a cyber-power index. Our profiling of cyber power would not ask how 'high' a state 'ranks' in relation to the cyber power other states, but rather whether any particular nation's overall national cyber power configuration is appropriately configured to serve that specific nation's unique national objectives, and in which respects it differs from any other state/s.

7. Conclusion

This paper conducted an exploratory analysis of cyber power in the African context.The analysis established the need for an expanded construct in the form of a cyber-power triad (comprising of offensive, defensive and developmental dimensions) and a tentative proposition was advanced on a notional ('abstract') level. This notional construct of a cyber-power triad is subject to, and will be concretised in, further research. It will culminate in the design and application of a qualitative framework for assessing cyber power of African countries. In sequential order, the sub-themes of the envisaged further research items are as follow:

- The African cyber power triad – concretising the developmental dimension.
- A high-level overview of an integrated framework for assessing cyber power in Africa.
- Cyber power in Africa: configuring capacity, challenges and constraints.
- A qualitative assessment of South Africa's cyber power.
- A comparative assessment of cyber power in the Southern African Development Community (SADC) region.
- Cyber power in sub-Saharan Africa: proposal on the cooperation agenda.

Albeit theoretical and academic, we expect qualitative assessments of this nature to be useful in informing African countries' leveraging of the asymmetric advantages that cyber space offer: offensively, defensively and developmental. Although not a panacea, such a synergetic wielding of cyber power is one of the keys indispensable to African countries addressing their substantial challenges and unlocking their vast potential.

References

African Union. 2015. *Flagship Projects of Agenda 2063*. As accessed on 16 January 2023 at https://au.int/agenda2063/flagship-projects.

African Union. 2020. *The digital transformation strategy for Africa (2020-2030)*. As accessed on 16 January 2023 at https://au.int/sites/default/files/documents/38507-doc-dts-english.pdf.

Allen, N. and La Lime, M. 2021. *How digital espionage tools exacerbate authoritarianism across Africa?* As accessed on 11 December 2022 at https://www.brookings.edu/techstream/how-digital-espionage-tools-exacerbate-authoritarianism-across-africa/

Allen, N. and van der Waag-Cowling N. 2021. *How African states can tackle state-backed cyber threats.* Brookings Techstream. As accessed on 09 December 2022 at https://www.brookings.edu/techstream/how-african-states-can-tackle-state-backed-cyber-threats/.

BAE Systems. 2022. *What is cyberpower*. Accessed on 27 November 2022 at https://www.baesystems.com/en/cybersecurity/feature/responsible-cyber-power.

Betz, D. and Stevens, T. 2011. *Cyberspace and the State. Toward a Strategy for Cyber-power*. London: International Institute for Strategic Studies.

Cavelty. M. D. 2018. 'Europe's cyber-power', *European Politics and Society*, vol. 19, nr 3: 304-320, DOI: 10.1080/23745118.2018.1430718

Dados, N. and Connel, R. 2012. *The Global South*. In Contexts, Vol. 11, No. 1, pp. 12-13. https://journals.sagepub.com/doi/pdf/10.1177/1536504212436479.

Devanny, J. 2021. *The Review and Responsible, Democratic Cyber Power*. As accessed on 11 December 2022 at https://www.kcl.ac.uk/the-review-and-responsible-democratic-cyber-power

Devanny, J., Goldoni, L and Medeiros, B. 2022. 'The rise of cyber power in Brazil' in *Revista Brasileira de Politica Internacional*, Vol 65, no 1. DOI: https://doi.org/10.1590/0034-7329202200113.

Duvenage, P.C. 2019. *A conceptual framework for cyber counterintelligence*, unpublished PhD thesis, University of Johannesburg, South Africa.

Çifci, H. 2022. *Comparison of National-Level Cybersecurity and Cyber Power Indices: A Conceptual Framework*. Istanbul: Istanbul Aydin University.

El Habti, H. 2022. *Why Africa's youth hold the key to its development potential*. World Economic Forum. As accessed on 16 January 2023 at https://www.weforum.org/agenda/2022/09/why-africa-youth-key-development-potential/

Fox, A.B. 1959. *The Power of Small States: Diplomacy in World War Two*. Cambridge: Cambridge University Press.

Gray, C.S. 2013. *Making Strategic Sense of Cyber Power: Why the Sky is not Falling*. Carlisle, PA: Strategic Studies Institute

Handler, S. 2021. *The 5×5—Cyber capacity and conflict in Africa*, Atlantic Council. Accessed on 29 November 2022 at https://www.atlanticcouncil.org/commentary/the-5x5-cyber-capacity-and-conflict-in-africa/

International Institute for Strategic Studies (IISSS). 2021. *Cyber Capabilities and National Power: Net Assessmen*t As accessed on 21 November 2022 at https://www.iiss.org/blogs/research-paper/2021/06/cyber-capabilities-national-power 2021)

International Telecommunication Union (ITU). 2021a. *Global Cybersecurity Index 2020*. Accessed on 27 October 2021 at https://www.itu.int/epublications/publication/global-cybersecurity-index-2020/en/.

International Telecommunication Union (ITU) (2021b), *Information and communication technology trends and developments in the Africa region. 2017-2020.* Accessed on 11 September 2022 at https://www.itu.int/pub/D-IND-DIG_TRENDS_AFR.01-2021.

Interpol. 2021. *The African Cyberthreat Assessment Report 2021*. Accessed on 09 December 2021 at https://www.interpol.int/content/download/16759/file/AfricanCyberthreatAssessment_ENGLISH.pdf.

Kegley, C.W. and Wittkopf, E.R. 1999*. World Politics: Trend and Transformation*. World Publishers: New York.

McCarthy, S. 2022. 'Africa needs to improve cybersecurity to boost investment', *World Economic Forum Opinion*. Accessed on 04 November 2022 at https://www.weforum.org/agenda/2022/08/africa-must-act-to-address-cybersecurity-threats.

Nye, J.S. 2010. *Cyber Power*. Cambridge, MA. Belfer Center for Science and International Affairs.

Odeh, L. E. 2010.' A Comparative Analysis of Global North and Global South Economies'. In *Journal of Sustainable Development in Africa* (Volume 12, No.3, 2010). Clarion, Pennsylvania. Clarion University of Pennsylvania.

Oxford Global Cybersecurity Centre: *Cybersecurity Capacity Maturity Model for Nations (CMM).* 2021. As accessed on 13 January 2022 at https://gcscc.ox.ac.uk/the-cmm.

Pagel, H., Ranke, K., Hempel, F. and Köhler, J. 2014. 'The Use of the Concept 'Global South'' in *Social Science & Humanities.* Berlin. Institut für Asien- & Afrikawissenschaften.

Segal, A. 2016. *The Hacked World Order: How Nations Fight, Trade, Maneuver, and Manipulate in the Digital Era.* New York, NY: Public Affairs.

Sheldon, J.B. 2011*. 'Deciphering Cyberpower: Strategic Purpose in Peace and War.'* In Strategic *Studies Quarterly*, Vol. 5 No. 2 (Summer 2011), pp. 95 – 112.

Spade, J.M. 2012. *Information as Power. China's Cyber Power and America's National Interest*. Carlisle, PA: US Army War College.

Starr, S.H. 2009. *Towards an Evolving Theory of Cyberpower*. Center for Technology and National Security Policy (CTNSP), National Defense University (NDU). United States.

Toje, A. 2010. *'The European Union as a Small Power.'* In *Journal of Common Market Studies*. Vol 49 (1).

Valeriano, B., Jensen, B and Maness, R.C. 2018. *Cyber Strategy: The Evolving Character of Power and Coercion*. New York: Oxford University Press.

van der Waag-Cowling, N.2019. 'Africa must harness the potential of Cyber Power and Digital Diplomacy' in Very, F (ed.) *Research Brief* 11/2019, Security Institute for Governance and Leadership in Africa, University of Stellenbosch.

van Niekerk, B. 2019. 'The Cyber Security Dilemma: A South African Perspective.' *Proceedings of the 14t^{h} International Conference on Cyber Warfare and Security,* University of Stellenbosch, South Africa .

Venables, A., Shaikh, S.A., and Shuttleworth, J. 2015. '*The Projection and Measurement of cyberpower*.' In Security *Journal, Vol. 30, no.* 3.

Voo, J., Hemani, I, and Cassidy, D. 2022. *National Cyber Power Index 2022*. Cambridge, MA. Belfer Centre for Science and International Affairs.

DPIA for Cloud-based Health Organizations in the Context of GDPR

Dimitra Georgiou and Costas Lambrinoudakis
Department of Digital Systems, University of Piraeus, Piraeus, Greece

dimitrageorgiou@ssl-unipi.gr
clam@unipi.gr

Abstract: The General Data Protection Regulation is the core instrument of the reformed legal framework for personal data protection in the European Union. The GDPR was put into effect on May 25, 2018, and requires assessing and conducting a Data Protection Impact Assessment for processing operations that are likely to result in a high risk to the rights and freedoms of natural persons, specifically using new technologies and considering the nature, scope, context, and purposes of the processing. Although GDPR does not precisely specify the types of processing activities for which a DPIA would be necessary, through the guidelines that it provides, the organization should conduct a DPIA, if there is large scale processing of health data. An example of this, is a Cloud-based Health Organization. Taking into account this parameter, that Cloud-based Health Organization processes personal data that could impact the freedoms and rights of a data subject under the GDPR and that the GDPR does not specify a DPIA process to follow, instead it allows organizations to use a framework that complements their existing processes, this paper presents the last two steps of a DPIA study for a Cloud-based Health Organization and provides guidelines on how to carry them out effectively. This study is part of a project for the compliance of Cloud-based Health Organizations with the General Data Protection Regulation 2016/679. For fulfilling the objectives of this study, the PIA-CNIL methodology is applied, which is in accordance with the data privacy impact assessment that has been described in ISO/IEC 29134. The main contribution of this work is the development of a guide that is designed to help Cloud-based Health organizations identify, analyze and reduce data protection risks in relation to their processing activities. More analytically, this research presents the risks that could be materialized by the data processing activities carried out by a Cloud-based Health Organization regarding its Processing Activities and could have an impact on the fundamental rights and freedoms of natural persons.

Keywords: cloud computing, security, privacy, data protection impact assessment, GDPR, healthcare systems

1. Introduction

The General Data Protection Regulation (European Union, 2016) is the most significant piece of privacy legislation to come into effect across Europe, affecting all sectors including healthcare. It is a common set of guidelines to control and to protect personal data and it brings significant changes to how organizations should manage and process personal data, the privacy risk assessments they conduct, and the privacy compliance programs they develop to mitigate the identified risks to the privacy of the data subjects. The GDPR applies to organizations located within the EU, as well as to organizations that offer services or monitor the behavior of EU residents regardless of the organization's location. Organizations serving every sector must respect the personal data they process, demonstrating compliance with the corresponding Regulation. The healthcare sector is among the most affected, for several reasons, including the problem of defining and dealing with the concept of personal data protection**.** And what does it mean? GDPR requires health organizations to take adequate measures to ensure the security of personal than with other types of data and a DPIA is mandatory for them. A DPIA is a risk technique mandated by the GDPR to enable organisations to address privacy concerns and ensure appropriate technical and organisational safeguards are addressed and built into new amendments of existing systems (David Wright, 2012). Under GDPR and Article 35, a DPIA is only mandatory where processing "...*is likely to result in a high risk to the rights and freedoms of natural persons*" (European Union, 2016). This includes automated decision-making, large-scale processing of special categories of data and systematic large-scale monitoring of public areas. As in the Cloud healthcarere domain all patients' data are considered "sensitive", it involves a large-scale processing volume of personal data and individuals are vulnerable. These characteristics meet three of the criteria described in (Chryssanthou et all, 2012) and therefore in the healthcare domain, data processing must be considered as possibly high-risk by default, thus implying that DPIA cannot be avoided. DPIAs are important tools for negating risk, and for demonstrating compliance with the GDPR (David Wright, 2011) Despite the recent attempts to provide tools to assist institutions to comply with the GDPR and as Article 35 (European Union, 2016) does not provide an explicit description for the DPIA and specifically for cloud-based healthcare systems (Kush Wadhwa & Rowena Rodrigues, 2013) (Makri Eleni-Laskarina et al, 2019).

To summarize, this paper apart from the introduction is organized as follows. Section 2 provides a description of DPIA steps of PIA-CNIL Methodology, Section 3 provides information for the Case Study of Cloud-Based Hospital, Section 4 presents the Privacy Impact Assessment related to the processing activities of 'Patients Monitoring Service' implemented by Hospital as Data Processor, Section 5 presents the Risk management

decision related to the processing activities of Patients Monitoring Service. Finally, Section 6 concludes the paper by outlining aspects for further research.

2. Description of DPIA steps of PIA-CNIL Methodology

For the analysis and the privacy impact assessment of personal data, the (French Data Protection Authority, 2018) has proposed PIA-CNIL methodology. Performed in principle by a data controller, the purpose of a PIA is to build and demonstrate the implementation of privacy protection principles so that data subjects retain control over their personal data. It is intended for data controllers who wish to demonstrate their compliance approach, as well as for product and service providers wishing to show that their solutions do not breach privacy thanks to a design that respects privacy satisfaction of Article 25 (European Union, 2016). More specifically, to carry out a PIA it is necessary to (Dimitra Georgiou and Costas Lambrinoudakis, 2021):

1. Define and describe the context of the processing of personal data under consideration and its stakes.
2. Identify existing or planned controls (procedural / technical / organisational) guaranteeing compliance with legal requirements and to treat privacy risks in a proportionate manner.
3. Assess privacy risks associated with data security and ensure they are properly treated.
4. Make the decision to validate the way it is planned to comply with privacy principles and treat the risks or review the preceding steps.

The following **Error! Reference source not found.** depicts the main steps of PIA-CNIL methodology (French Data Protection Authority, 2018) and mentions the step 3 and step 4 that will be analysed in this paper. Furthermore, the objectives of all the steps of PIA- CNIL methodology are presented in the following paragraphs.

Figure 1: Steps of PIA-CNIL methodology (French Data Protection Authority, 2018)

The **first step – Context of personal data processing** aims at the definition of the outline of the processing of personal data, such as the categories of the processed personal data, their ways of processing, the purpose of processing, the personal data supporting assets, the data subjects, etc. This step contains the definition of the valuable for the under-examination system Assets:

- The description of the purpose of processing(s) of personal data.
- The identification of Data Controller and any Data Processor(s).
- The identification of the categories of personal data and their recipients.
- The identification of the retention period of personal data.
- The description of the processes and personal data supporting assets for the entire personal data life cycle (from collection to erasure).

The objective of the **second step – Controls** is to build a system that ensures compliance with privacy protection principles. This step requires compliance (and thorough documentation of the way to achieve it) with the following legal requirements (obligatory):

- Purpose of personal data processing: Specified, explicit and legitimate purpose.
- Data Minimisation: limiting the amount of personal data to what is strictly necessary.
- Quality of data: preserving the quality of personal data, accurate and kept up-to-date.
- Retention periods: period needed to achieve the purposes, in the absence of another legal obligation imposing a longer retention period.
- Information: respect for data subjects' right to information.
- Consent: obtaining the consent of the data subjects or existence of another legal basis justifying the processing of personal data.
- Right to object: respect for the data subjects' right of opposition.
- Right of access: respect for the data subjects' right to access their data.
- Right to rectification: respect for the data subjects' right to correct their data and erase them.
- Transfers: compliance with obligations relating to transfer of data outside the European Union

Moreover, the existing controls are identified or determined:

- Organisational controls: organisation policy, risk management, project management, incident management, supervision, etc.
- Logical security controls: anonymisation, encryption, backups, data partitioning, logical access control, etc.
- Physical security controls: physical access control, security of hardware, protection against non-human risk sources, etc.

The objective of the **third step** Risks (Potential privacy breaches) of PIA-CNIL is to gain a good understanding of the causes of risks, the threats against privacy, as well as the impact of their potential realisation. In this step, the following should be defined:

- Sources of Risks
    - Identification of the relevant risk sources in the specific context under consideration
    - Description of the capabilities of risk sources
- Feared events
    - For each feared event (illegitimate access to personal data, unwanted change of personal data, and disappearance of personal data):
    - Determination of the potential impacts on the data subjects' privacy if it occurred.
    - Estimation of its severity, depending especially on the prejudicial effect of the potential impacts
- Threats
    - Identification of threats to personal data supporting assets that could lead to each feared event.
    - For each identified threat:
        - Selection of the risk sources that could cause it.
        - Estimation of its likelihood, particularly depending on the level of vulnerabilities of personal data supporting assets, the level of capabilities of the risk sources to exploit them and the controls likely to modify them.
- Risks
    - Determination of the risk level:
        - Its severity equals to that of the feared event concerned by the risk.
        - Its likelihood equals the highest likelihood value of the threats associated with the feared event.

The objective of the **fourth step - Risk management decisions** is the review of the results of the preceding steps, the evaluation of the risk level and the already existing controls, and the determination whether or not they are acceptable. In case modifications are needed, an action plan is developed for the improvement of this state. In this step, the already existing controls are evaluated for the satisfaction of legal requirements and decisions are made whether existing controls are satisfactory. When not, an action plan is prepared and validated.

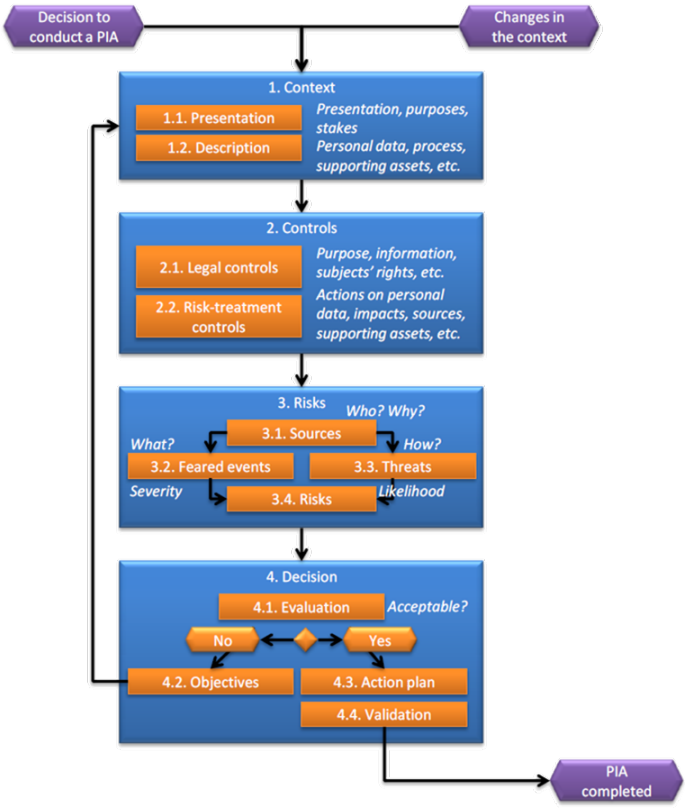


Figure 2: Detailed approach for carrying out PIA-CNIL Methodology (French Data Protection Authority, 2018)

3. Case Study: A Cloud-Based Hospital -Theoretical Background

This document primarily refers to the case study of a Cloud-based Hospital, however its results are applicable to all entities as far as organisational and operational issues are concerned. Furthermore, different Organisations can use the approach used in this case study as guidance of the information that needs to be included in a DPIA, making adaptations where necessary depending on the circumstances of the case. Information regarding the on-premises ICT infrastructure is only applicable to this case study. In this regard, we indicate the steps of guidelines in designing and testing a DPIA process that is generic enough to be used in all possible application areas, but also able to consider the specifics of each area. For fulfilling the objectives of this study, PIA-CNIL methodology (French Data Protection Authority, 2018) is applied. This paper presents the Steps 3 and 4 of the DPIA Methodology, risks that could be materialised by the data processing activities carried out by Hospital regarding its Processing Activities and could have an impact on the fundamental rights and freedoms of natural persons. These Processing Activities will be implemented directly by Hospital as Data Processor. The risks for the fundamental rights and freedoms of natural persons that will be assessed in the context of the impact assessment in this paper are related with *illegitimate access to data, unwanted modification of data and data disappearance*. Since the analysis that has been conducted in paper (Dimitra Georgiou and Costas Lambrinoudakis, 2021) has revealed that this specific processing activity PP4: Patients Monitoring Service involves special categories of personal data on a large scale which could be characterised as a high risk processing and a Gap Analysis in relation to the GDPR requirements for this processing activity has been conducted, it is necessary to perform a privacy impact assessment study, as the hospital carries out the processing of special categories of personal data on a large-scale even though this is an obligation of the Data Controller and not of the Hospital.

Impact assessment aims at the protection of personal data, as well as the protection of elements that support their processing and are recognised as Assets. The value of such assets is equal to the Impact brought upon by a possible violation of individuals' privacy. A Feared Event is the illegitimate access to personal data, unwanted modification of personal data, as well as the data disappearance. The violation of information systems needs the existence of Vulnerability and the appearance of a relevant Threat coming from a Risk source. Summarising, we note that a Threat exploits a vulnerability of an Information System and can have as a result an incident of data protection breach, inflicting some Impact on data subjects (**Figure** 33). The risk level is estimated in terms

of severity, which represents the magnitude of a risk. It essentially depends on the prejudicial effect of the potential impacts, and likelihood, which represents the possibility for a risk to occur. It essentially depends on the level of vulnerabilities of the supporting assets facing threats and the level of capabilities of the risk sources to exploit them (**Figure 3**).

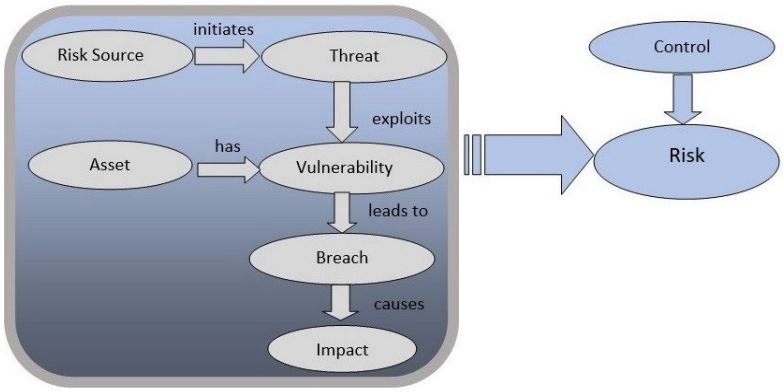


Figure 3: Conceptual Framework of Impact Assessment

4. Assessment of Privacy Risks related to the processing of "Patients Monitoring Service" (PP4)

4.1 Introduction

The following paragraphs describe the impact assessment for the potential violation of the privacy of the Cloud hospital's personal data (Georgiou and Lambrinoudakis, 2021) involved in the processing activities for the purpose of "*Patients Monitoring Service (PP4)".* This section analyses the personal data processed by hospital and the level of risk by determining the severity of the risk, which is the same as the severity of the data breach incident (feared event) that can be caused by the specific risk and by the likelihood. According to the PIA-CNIL methodology (French Data Protection Authority, 2018) feared events may have three different impacts if they occur (**Table** 1) causing data breaches. The severity of risk and the likelihood of a risk occurring are scaled in levels as presented later.

Table 2 contains the scales and rules for estimating severity, in terms of the extent of potential impacts on data subjects. Section 4.2 describes the processing activities related with platform and the purpose of Patients Monitoring Service, implemented by Hospital as Data Processor. All three categories of risks are analysed, as it is estimated that they can be materialised by the corresponding processing activities (Georgiou and Lambrinoudakis, 2021).

Table 1: Typology of the outcomes of feared events (French Data Protection Authority, 2018)

Feared events	Types of outcomes	Description
Illegitimate access to data	None	The data are seen by people who do not need to know them, though these people do not use them.
	Storage	The data are copied and saved to another location without being further used.
	Redistribution	The data are disseminated more than necessary and beyond the control of the data subjects.
	Use	The data are used for purposes other than those planned and/or in an unfair manner or correlated with other information relating to the data subjects.
Unwanted modification of data	Malfunction	The data are modified into valid or invalid data, which will not be used correctly, the processing liable to cause errors, malfunctions, or no longer provide the expected service.
	Use	The data are modified in other valid data, such that the processing operations have been or could be misused.
Disappearance of data	Malfunction	The data are missing for personal data processing, which generates errors, malfunctions, or provides a different service than the one expected
	Blockage	The data are missing for personal data processing which can no longer provide the expected service

Table 2: Scales and rules for estimating severity (French Data Protection Authority, 2018)

Levels	Generic description of impacts
Negligible	Data subjects either will not be affected or may encounter a few inconveniences, which they will overcome without any problem.
Limited	Data subjects may encounter significant inconveniences, which they will be able to overcome despite a few difficulties.
Significant	Data subjects may encounter significant consequences, which they should be able to overcome albeit with real and serious difficulties.
Maximum	Data subjects may encounter significant, or even irreversible, consequences, which they may not overcome.

The aforementioned scale can be used to estimate the likelihood of threats (French Data Protection Authority, 2018):

1. Negligible: It does not seem possible for the selected risk sources to materialise the threat by exploiting the properties of supporting assets.
2. Limited: It seems difficult for the selected risk sources to materialise the threat
3. Significant: It seems possible for the selected risk sources to materialise the threat
4. Maximum: It seems extremely easy for the selected risk sources to materialise the threat.

4.2 Privacy Impact Assessment related to the processing activities of "Patients Monitoring Service"

Based on the requirements of the GDPR (Georgiou and Lambrinoudakis, 2020)and taking into account the criticality of the data processed by Hospital for the *Patients Monitoring* that Hospital implements as Data Processor (Georgiou and Lambrinoudakis, 2021)**,** we consider that it is necessary to perform a PIA study, as Hospital carries out processing of special categories of personal data on a large scale, which could be characterised as a high risk processing, according to the provisions of the Article 35 of GDPR (European Union, 2016).

The Hospital care platform is used to enable data processing for the purpose of Patients Monitoring. Hospital conducts processing of personal data and special categories of data that are hosted at the databases of the Hospital care platform. More specifically, Hospital supports through the platform staff or patients by managing therapies for both inpatients and outpatients. Hospital infusion devices communicate with Hospital care platform and transmit device data, which are stored and presented under the treatment context. The risks for the fundamental rights and freedoms of natural persons that are assessed in the context of the impact assessment are related with *illegitimate access to data, unwanted modification of data and data disappearance.*

4.2.1 Risk category: Illegitimate access to data

The risk of unauthorised access to data that Hospital care platform maintains can appear in the processing activities for patients monitoring due to various threats (Figure 3). This section discusses these threats grouped according to their nature or to the asset that they target and include masquerading of identity, unauthorised use of an application, threats during data transmission and misuse of physical resources of Hospital. The analysis of the impact of each threat is carried out in accordance with Typology of the outcomes of feared events (Table 1) and considers the potential for unauthorised storage, redistribution, or use of personal and sensitive data. Next, the components of risk per threat are analysed in the following parts:

Threat: Masquerading of User Identity
• **Masquerading of User Identity by Insiders** – The threat covers attempts by authorised users to gain access to information to which they have not been granted access. • **Masquerading of User Identity by Contracted Service Providers –** The threat covers attempts by people working for a contracted service provider to obtain unauthorised access to information by using an authorised person. • **Masquerading of User Identity by Outsiders** – The threat covers attempts by outsiders to obtain unauthorised access to information by posing as an authorised user.

Figure 4: Threat Masquerading of User Identity

Impact: Inappropriate use of patients' personal data, due to masquerading of users may result in the use of

sensitive data of platform for purposes other than the purpose originally set. *Severity*: Severity from threats of masquerading vary depending on the occurrence of the threat, but in its maximum appearance it leads to a maximum impact and is valued at level 4. *Likelihood:* Hospital implements some technical controls specific to protecting the digital identity of Hospital care users. Additionally, there has been no similar occurrence in the past, but incidents of stolen credentials are common. Based on this, it is estimated that the probability of the threat is negligible and is valued at level 1. *Risk sources:* employees of Hospital, former Hospital employees, external attackers, and any other Hospital care user, Malicious code of unknown origin. *Controls:* For the estimation of the impact, the implemented by Hospital controls have been considered.

Threat: Unauthorised Use of an Application
This Threat covers attempts of use of resources of technical aspects, in a non-compatible way to the provided authorisations and rights.

Figure 5: Threat Unauthorized Use of an Application

Impact: Incorrect use of data due to unauthorised use of an application may prevent the Hospital's provision of Hospital care services while the data may be used for purposes other than the purpose originally set. *Severity:* The severity of unauthorised access leads to a maximum impact and is valued at level 4. *Likelihood:* Hospital, imposed by hospital's security policy, takes high technical and organisational controls to prevent unauthorised access, such as strong password policy, multifactor authentication, https. Additionally, there has been no similar occurrence in the past. Based on this, it is estimated that the probability of the threat is limited and is valued at level 1. *Risk sources:* employees and former hospital employees, malicious code of unknown origin. *Controls:* For the estimation of the impact, the implemented by Hospital controls have been considered.

Threat: Threats during data transmission
• **Communication Interception** – This threat covers: passive interception and traffic monitoring. • **Accidental Mis-routing** – The threat covers the possibility that information might be delivered to an incorrect address.

Figure 6: Threat during the data transmission

*Impact***:** Attacks aimed at interfering with electronic communications, traffic monitoring or accidental misrouting can lead to leakage of patients' data. *Severity***:** Data transmission threats may lead to a significant impact and is valued at level 3. *Likelihood:* Hospital implements technical and organisational controls to prevent unauthorised access, such as https. Additionally, there has been no similar occurrence in the past. Based on this, it is estimated that the probability of the threat is negligible and is valued at level 1. *Risk sources:* employees and agents of Hospital who have access to the platform, former employees, external attackers, or any other user, malicious code of unknown origin. *Controls:* For the estimation of the impact, the implemented by hospital controls have been considered.

Threat: Misuse of physical resources
• **Theft (by Insiders)** – This threat would include anybody who had a legitimate reason to be working in the building. • **Theft (by Outsiders)** –Theft of any resources, as physical assets. It relates to thefts by outsiders. • **Terrorism** – Covers acts by extremist groups wishing to cause damage to the work of the organisation, or harm people

Figure 7: Threat Misuse of physical resources

Impact: Unauthorised access to data due to misuse of resources may prevent the hospital's provision of services and lead to significant impacts for both hospital and for the patients. *Severity:* Severity from misuse of physical resources threats leads to a significant impact and is valued at level 3. *Likelihood:* From hospital's point of view, the workspace is well monitored and there is access control system. Combining this with the absence of a past incident, it is estimated that the probability of the threat is negligible and valued at level 1. *Risk sources:* employees of hospital, former employees, and any other user, who had access to platform. *Controls:* For the estimation of the impact, the implemented by hospital controls have been considered. The list of implemented controls, will be presented in our new paper.

4.2.2 Risk category: Unwanted modification of data

The risk of unintended modification to Hospital data can be realised in the processing activities based upon various threats (Figure 3). This section addresses these threats grouped according to their nature or according to the asset they exploit, including masquerade, damage to hardware, and damage to software. The components of risk per threat are analysed in next paragraphs:

Threat: Masquerading of User Identity
• **Masquerading of User Identity by Insiders** – The threat covers attempts by authorised users to gain access to information to which they have not been granted access. • **Masquerading of User Identity by Contracted Service Providers –** The threat covers attempts by people working for a contracted service provider to obtain unauthorised access to information by using an authorised person. • **Masquerading of User Identity by Outsiders** – The threat covers attempts by outsiders to obtain unauthorised access to information by posing as an authorised user.

Figure 8: Threat Masquerading of user Identity

Impact: The masquerading of a user's identity can lead to unwanted modification of patients' data. This action can lead to a maximum impact on patients as the modification of information, can cause harm to the patient that may even lead to death. *Severity:* Based on the above, severity from threats vary depending on the occurrence of the threat, but in its maximum appearance it leads to a maximum impact and is valued at level 4. *Likelihood:* Hospital implements specific technical controls to protect the digital identity of Hospital care users. Additionally, there has been no similar occurrence in the past, but incidents of stolen credentials are common. Based on this, it is estimated that the probability of the threat is negligible and is valued at level 1. *Risk sources:* employees of Hospital, former Hospital employees, external attackers, malicious code of unknown origin. *Controls:* For the estimation of the impact, the implemented by Hospital controls have been considered.

Threat: Hardware Malfunction
Operations Error – The threat covers the possibility that the people responsible for operating the Host system might make mistakes when carrying out their work. **Hardware Maintenance Error** –The threat covers the possibility that those people responsible for maintaining the hardware might make mistakes when carrying out their work. **Staff Shortage** –The threat covers the possibility of the absence of key personnel for whatever reason and the ease with which they could be replaced.

Figure 9: Threat Hardware Malfunction

Impact: Insufficient hardware maintenance or operating mistakes on the hosting platform of hospital, may lead to availability or integrity loss of provided services. *Severity***:** severity from threats vary depending on the occurrence of the threat, but in its maximum appearance it leads to limited impact and is valued at level 2. *Likelihood:* Given to the contractual obligations and SLAs available, the possibility of hardware maintenance/operations errors, or insufficient support, is low. Based on this, it is estimated that the probability of the threat is valued at level 1. *Risk sources***:** Employees of hospital, incorrectly configured legal binding documents. *Controls:* For the estimation of the impact, it has been considered that hospital exhibits several security certifications and thus it has been assumed that the appropriate controls are in place.

Threat: Software Malfunction
• **Unauthorised Use of an Application** – The threat covers attempts of use of resources of technical aspects, in a non-compatible way to the provided authorisations and rights. • **Introduction of Damaging or Disruptive Software** – The threat covers: any forms of malicious software. • **Embedding of Malicious Code** – The threat covers: e-mail viruses, hostile mobile code. • **Software Maintenance Error** – The threat covers the possibility that those people or organisations responsible for maintaining software might make mistakes when carrying out their work.

Figure 10: Threat Software Malfunction

Impact: Damage to the software may result in modification of patients' data. This action can lead to a limited impact on data subjects as it can cause consequences for them. *Severity***:** Severity from software malfunction threats leads to a limited impact and is valued at level 2. *Likelihood:* Hospital takes appropriate technical and organisational controls to protect its software. Additionally, there has been no similar occurrence in the past. Based on this, it is estimated that the probability of the threat is negligible and is valued at level 1. *Risk sources:* employees who are responsible for handling Hospital systems and have access to the platform, external

attackers, malicious code of unknown origin. *Controls:* For the estimation of the impact, the implemented by Hospital controls have been considered.

4.2.3 Risk category: Data disappearance

This section discusses these threats grouped according to their nature or according to the asset that they exploit and includes masquerading, technical failure, application software failure, communication breaches, malfunction to physical resources of Hospital. The analysis of the impact of each threat is conducted according to **Error! Reference source not found.**(**Table** 1) and takes into account the possible malfunctioning processing and error causing through the processing of data or the possible loss of personal and sensitive data when they can no longer provide the expected service to data subjects. Next, the components of risk per threat are analysed.

Threat: Masquerading of User Identity
• **Masquerading of User Identity by Insiders** – The threat covers attempts by authorised users to gain access to information to which they have not been granted access. • **Masquerading of User Identity by Contracted Service Providers –** The threat covers attempts by people working for a contracted service provider to obtain unauthorised access to information by using an authorised person. • **Masquerading of User Identity by Outsiders** – The threat covers attempts by outsiders to obtain unauthorised access to information by posing as an authorised user.

Figure 11: Threat Masquerading of User Identity

Impact: The masquerading of any users' platform identity from internal or external attackers can lead to the deletion of the data. This action can lead to maximum impact on patient, as the unavailability of certain data, that may render a doctor incapable of providing appropriate treatment. *Severity:* Severity from threats vary depending on the occurrence of the threat, but in its maximum appearance it leads to a maximum impact and is valued at level 4. *Likelihood:* Hospital takes some specific technical controls to protect the digital identity of users. Additionally, there has been no similar occurrence in the past, but incidents of stolen credentials are common. Based on this, the probability of the threat is negligible and is valued at level 1. *Risk sources:* Malicious code of unknown origin. *Controls:* For the estimation of the impact, the implemented by Hospital controls have been considered.

Threat: Technical Failure
• **Technical Failure of Host** – This threat covers failures of the CPU or other hardware items. • **Technical Failure of Storage Facility** – This threat covers disk crashes and disk failures. • **Technical Failure of Network Distribution Component** – This threat covers failure of the network distribution component, • **Technical Failure of Network Gateway** – This threat covers failure of servers or network operation. • **Air Conditioning Failure** – The threat covers the possibility that work may have to be suspended because temperatures in the location fall outside of acceptable parameters.

Figure 12: Threat: Technical Failure

Impact: Possible technical failure of network components, can lead to loss of availability of provided services. This action can lead to a significant impact on patients as the integrity of their data is at stake. *Severity:* severity from threats of technical failure vary depending on the occurrence of the threat, but in its maximum appearance it leads to a significant impact and is valued at level 3. *Likelihood:* Hospital offers redundant components and thus availability of the provided service of the users may be lost. Based on this, it is estimated that the probability of the threat is limited and is valued at level 2. *Risk sources:* hospital's relevant hardware components in infrastructure. *Controls:* For the estimation of the impact, it has been considered that hospital exhibits several security certifications and thus it has been assumed that the appropriate controls are in place.

Threat: Application Software Failure
• **System and Network Software Failure** – The threat covers the possibility that the system or network software might fail. • **Application Software Failure** – The threat covers the possibility of errors being contained in application programs.

Figure 13: Threat Application Software

Impact: Software failure, can lead to the deletion of personal and sensitive data hosted on it. Appearance of the above threat is expected to have a limited impact on data subjects as, since appropriate controls, like back-up are available. *Severity:* Based on the above, severity from threats vary depending on the occurrence of the threat, but in its maximum appearance it leads to a negligible impact and is valued at level 1. *Likelihood:* It is estimated that the realisation of the threat is random, its probability is small, so it is valued at level 1. *Risk sources:* Application programming errors and bugs. *Controls:* For the estimation of the impact, the implemented by Hospital controls have been considered.

Threat: Communication breaches
• **Communication Infiltration** – The threat covers Hacking into a system, masquerading as a server, masquerading as an existing user of an e-commerce application, masquerading as a new user of an e-commerce application, denial of service, flaming attacks, spamming. • **Communication Manipulation** – The threat covers active interception, insertion of false messages, deliberate delivery out of sequence, deliberate delay of delivery, deliberate misrouting, if an attacker can force a message to be sent via a hostile host, the attacker may be in a position to intercept, alter and the forward the message. • **Communication Failure** – This threat covers: Unavailability of Service Provider, failure of data link, non-delivery of message, accidental delivery out of sequence, accidental delay in delivery, accidental denial of service.

Figure 14: Threat Communication Breaches

Impact: This category of threats includes attacks aimed at editing, inserting, or deleting data. *Severity:* The threat severity from threats to communications differs depending on the occurrence of the threat but in its maximum appearance it leads to a significant impact and is valued by a level 3. *Likelihood:* The hospital takes technical and organisational controls to protect communications and its network, such as firewall and https. In addition, no attacker has been documented with an important incentive to carry out such an attack. Based on this, it is estimated that the probability of the threat is limited and valued at level 2. *Risk sources:* former Hospital employees who had access to platform, malicious code of unknown origin. *Controls:* For the estimation of the impact, the implemented by Hospital controls have been considered.

Threat: Malfunction to physical resources
• **Power failure** – The threat covers the possibility that the power supply to the building may fail. • **Fire** – The threat covers the possibility of fire affecting any of the physical assets that make up a system. • **Water Damage** – The threat covers the possibility of water affecting any of the physical assets that make up a system. • **Natural disaster** – The threat of natural disaster covers the possibility of either a natural event, or manmade (such as traffic accidents), causing physical damage to the location or surrounding area. • **Wilful damage – vandalism (by Insiders)** – The threat covers acts of vandalism and other cases where physical damage is caused to IT systems or their supporting environment by people who have been granted access to the building. • **Wilful damage – vandalism (by Outsiders)** – The threat covers acts of vandalism and other cases where physical damage is caused to IT systems or their supporting environment by people who have not been granted access to the building. • **Terrorism** –The threat covers acts by extremist groups wishing to cause damage or disruption to the work of the organisation, or harm people working for the organisation. Types of terrorist attack include letter bombs and car bombs.

Figure 15: Threat Malfunction to physical resources

Impact: Possible failure can lead to loss of availability of provided services. *Severity:* severity from the above threats vary depending on the occurrence of the threat, but in its maximum appearance it leads to a significant impact and is valued at level 3. *Likelihood:* Hospital takes particularly technical and organisational controls to protect from malfunction to physical resources. Based on this, it is estimated that the probability of the threat is limited and is valued at level 1. *Risk sources:* Employees of hospital, Former employee's terrorist organisations, fire, flood, power failure. *Controls:* For the estimation of the impact, hospital exhibits security certifications and thus it has been assumed that the appropriate controls are in place.

4.2.6 Overview of risks

The overall results of the risk assessment per risk category are outlined in the following Figures:

Risk category: Illegitimate access to data		
Threat	**Severity**	**Likelihood**
➢ Masquerading of User Identity	Maximum, Level: 4	Negligible, Level: 1
➢ Unauthorised Use of an Application	Maximum, Level: 4	Negligible, Level: 1
➢ Threats during data transmission	Significant, Level: 3	Negligible, Level: 1
➢ Misuse of physical resources	Significant, Level: 3	Negligible, Level: 1
MEAN ASSESSMENT VALUE	**3,5≈4 Maximum Level: 4**	**1 Negligible Level: 1**
GRAPHIC ANALYSIS		
Severity	(Undefined) Negligible Limited Significant	
Likelihood	(Undefined) Negligible Limited Significant	

Figure 16: Results of the Risk assessment for the illegitimate access to data

Risk category: Unwanted modification of data		
Threat	**Severity**	**Likelihood**
➢ Masquerading	Maximum, Level: 4	Negligible, Level: 1
➢ Hardware Malfunction	Significant, Level: 2,5≈3	Negligible, Level: 1
➢ Software Malfunction	Limited, Level: 2	Negligible, Level: 1
MEAN ASSESSMENT VALUE	**2,8≈3 Significant Level: 3**	**1 Negligible Level: 1**
GRAPHIC ANALYSIS		
Severity	(Undefined) Negligible Limited Significant Maximum	
Likelihood	(Undefined) Negligible Limited Significant Maximum	

Figure 17: Results of the Risk assessment for the Unwanted modification of data

Risk category: Data disappearance		
Threat	**Severity**	**Likelihood**
➢ Masquerading	Maximum, Level: 4	Negligible, Level: 1
➢ Technical failure	Significant, Level:3	Limited, Level:2
➢ Application Software Failure	Limited, Level: 2	Negligible, Level: 1
➢ Communications breaches	Significant, Level: 3	Limited, Level: 2
➢ Malfunction to physical resources	Significant, Level:3	Negligible, Level: 1
MEAN ASSESSMENT VALUE	**3 Significant Level: 3**	**1,3≈1 Negligible Level: 1**
GRAPHIC ANALYSIS		
Severity	(Undefined) Negligible Limited Significant Maximum	
Likelihood	(Undefined) Negligible Limited Significant Maximum	

Figure 18: Results of the Risk assessment for the Data disappearance

5. Risk management decision related to the activities of Patients Monitoring Service

The objective of the Step 4 is the review of the results of the preceding steps, the evaluation of the risk level and the already existing controls and the determination whether they are acceptable. In case modifications are needed, an action plan is developed for the improvement of this state. In this step, the already existing controls are evaluated for the satisfaction of legal requirements and decisions are made whether existing controls are satisfactory. When not, an action plan is prepared and validated. A Risk mapping enables graphical identification of the risk, which in turn enables the carrier to identify parts of the infrastructure and its functions, which require corrective actions to address the risks arising for the protection of rights and freedoms of data subjects. Based on the identification of the risks in this situation, targeted corrective actions are chosen seeking to reduce the

impact of a risk or the likelihood of a threat. The corrective actions as well as the intended risk minimisation that follows the implementation of these actions will be presented in a new paper.

6. Conclusions and Further Research

Compliance with GDPR requirements is a challenge for the entire health community, as technological advances expand the frontiers of areas such as Cloud Computing storage. Complex and large-scale data processing activities in the health sector require careful planning and execution. An important tool to ensure that all relevant stakeholders in an organization assess GDPR requirements is the DPIA. In this paper we described a methodology specific for the Cloud-based Health Information Systems able to support the risk assessment and to perform a DPIA. The main contribution of this work is the development of a guide that is designed to help Cloud-based Health Organizations identify the information system threats (cclassification of threats) per risk category and reduce data protection risks in relation to their processing activities. The methodology was successfully applied in a real environment The risks examined in this paper are related with illegitimate access to data, unwanted modification of data and data disappearance. For the processing activities of Patients Monitoring Service, implemented by Hospital as Data Processor, the threats that may lead to the aforementioned risks and have the maximum likelihood to occur were: Masquerading of User Identity, Misuse of physical resources, Technical failure, Communication breaches, Malfunction to physical resources. From the aforementioned threats, the ones with the maximum severity to the rights and freedoms of natural persons were: Masquerading of User Identity (maximum severity), Misuse of physical resources (significant severity), Communication breaches (significant severity). The risk illegitimate access to data is assessed with a maximum level of severity and limited likelihood of occurring. The risks unwanted modification of data and data disappearance are assessed with a significant level of severity and limited likelihood of occurring. As the factors determining the risk level are the severity of the impact and likelihood of occurring. In a next paper, we will present the controls that will be chosen to be applied so as to demonstrate compliance with the GDPR and to offer the appropriate level of protection.

References

Chryssanthou et all, 2012. Hospital Information Systems Replacement and Healthcare Quality. *International Journal of Reliable and Quality E-Healthcare (IJRQEH),* 1(3), p. 12.

David Wright, 2011. Should privacy impact assessments be mandatory?. *Communications of the ACM,* August, 54(8), p. 121–131.

David Wright, P. H., 2012. Privacy Impact Assesment. In: P. H. David Wright, ed. *Law, Governance and Technology Series.* Dordrecht: Springer, pp. 3-32.

European Union, 2016. REGULATION (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Reg. *Official Journal of the European Union.*

French Data Protection Authority, 2018. *Commission Nationale de l'Informatique et des Libertés.* [Online] Available at: https://www.cnil.fr/en/privacy-impact-assessment-pia [Accessed 1 09 2022].

Georgiou and Lambrinoudakis, 2020. Compatibility of a Security Policy for a Cloud-Based Healthcare System with the EU General Data Protection Regulation (GDPR). *Information,* 11(12)(586), p. 19.

Georgiou and Lambrinoudakis, 2021. Data Protection Impact Assessment (DPIA) for Cloud-Based Health Organizations. *Future Internet,* 7 March, 13(3)(66), p. 12.

Kush Wadhwa & Rowena Rodrigues, 2013. Evaluating privacy impact assessments, 26:1-2,. *Innovation: The European Journal of Social Science Research,,* 21 March, 26(1-2), pp. 161-180.

Makri Eleni-Laskarina, G. Z. L. C., 2019. A proposed privacy impact assessment method using metrics based on organizational characteristics. In: *Computer Security: ESORICS 2019 International Workshops, CyberICPS, SECPRE, SPOSE, and ADIoT.* Luxembourg: s.n., pp. 122-139.

Radiograph Manufacturer and Model Identification Using Deep-RSI

Farid Ghareh Mohammadi[1,2] and Ronnie Sebro[1,2]
[1]Department of Radiology, Mayo Clinic, Jacksonville, Fl, USA
[2]Center for Augmented Intelligence, Mayo Clinic, Jacksonville, Fl, USA

Gharehmohammadi.farid@mayo.edu
Sebro.ronnie@mayo.edu

Abstract: Malware attacks of healthcare institutions are simultaneously becoming more common and more sophisticated. Artificial intelligence (AI) has resulted in the ability to rapidly alter or generate false images, advancing the ease of forgery of digital images. Digital image manipulation and substitution of radiographs are major threats to healthcare institutions because these altered images may affect patient care. Identifying the source (manufacturer, model) of radiology images is one method of validating the origin of radiology images in a healthcare system. In a previous study, researchers demonstrated that features from magnetic resonance imaging (MRI) could be used to trace and authenticate the source of the MRI images. We previously developed and tested the Deep learning for Radiograph Source Identification (Deep-RSI) approach for source identification of radiographs obtained of the upper extremities (hands, wrists, forearms, elbows, and shoulders). In this research, we present an empirical and quantitative investigation using deep learning to validate the source of digital radiographic images of the lower extremities (knees, legs, ankles, and feet). A convolutional neural network (CNN) is employed to extract features, which are then followed by three fully connected layers (FCNN). To ensure that our proposed method is a content-free approach, we added a new layer before the CNN to extract the initial content-free pixels and train the features using the CNN and FCNN layers. This proposed approach was used to identify the source of each digital image of a lower extremity. Adult patients of both sexes who had radiographs of the lower extremities at Mayo Clinic between 01/01/2010 and 12/31/2021 were evaluated. The data was randomly split by patient into training/validation and test datasets. There were 9 radiographic machine models and 6 manufacturers. Deep-RSI had an accuracy of 99.00% (AUC= 0.99) and 97.00% (AUC=0.94) for detecting the manufacturer and model of the radiographic machine for radiographs of the feet respectively, confirming that forensic evaluation of radiographs can be performed. This is the first medical forensics examination of this type to identify and confirm the source origins for radiographs of the lower extremities. This technique may be helpful to detect radiology malware attacks and scientific fraud.

Key words: Deep Learning, Medical Forensics, Lower Extremities, Radiology, Source Identifications

1. Introduction

The main purpose of forgery of an image is to hide or alter details in the images so that the images can be used to falsely convey different information (Jaiswal & Srivastava, 2022). Artificial Intelligence (AI) is currently used in digital image manipulation, which has become a major issue in a variety of industries, including social media, medical imaging, digital forensics and scientific publications. As AI technology advances it is more difficult for a professional to discern a real from fake image. This increases of the likelihood of a successful digital image forgery. Digital image manipulation and computer-generated fake radiographs are extreme threats to healthcare institutions because these altered radiographs may adversely affect patient care. This will have a negative impact on patients and respected institutions.

Generative Adversarial Networks (GANs) (Goodfellow et al, 2020) are a type of generative deep learning model that generate computer-generated fake images and text data (Deepfakes). Hackers use these advanced tools to increase the likelihood of seamless forging of digital images. Deepfakes have arisen as a severe threat to the security of healthcare institutions and is a significant threat particularly in radiology because fake digital radiographs are difficult to detect. Currently radiologists interpret radiographs without confirming of the source of the images or verifying that the radiographs are not fakes. However, if these radiographs are altered or manipulated, radiologists may be unable to detect real from fake radiographs. Altering or replacing radiographs poses serious risks to healthcare institutions because these radiographs are used for diagnosis. This can delay patient care and affect patient prognosis. Validating the radiographs through detecting the source (manufacturer and model) of the radiographs is an important step in finding the solution to this important issue. There are a few proposed solutions employing AI to detect this forgery in medical images which are limited to (Fang et al, 2020), (Ghareh Mohammadi & Sebro, 2022), so we propose expanding this fraud detection methodology using a different dataset (radiographs of the lower extremities) by utilizing the source identification solution. In this paper, we propose to leverage AI to predict the radiographs sources (manufacturer and model of radiographic machine) to detect forgery of digital radiographs.

2. Materials and Methods

The study protocol was reviewed and approved by the Mayo Clinic Institutional Review Board (IRB). Data collection was in accordance with the Health Insurance Portability and Accountability Act (HIPAA). We followed the ethical principles for medical research involving human subjects, including research on identifiable human material and data according to Declaration of Helsinki.

Patients

The study patients were individuals who had radiographs of the knees, legs, ankles, and feet at Mayo Clinic between 01/01/2010 and 12/31/2021. We randomly split the patients into train/validation (n=80%) and test(n=20%) datasets, so that the dataset partitions are independent.

Radiographic machines

Radiographs were taken using General Electric (GE) healthcare (General Electric, Waukesha WI), Canon (Tustin, CA), Carestream Health (Rochester, NY), Philips Medical Systems (Amsterdam, Netherlands), and Siemens Healthineers (Erlangen, Germany) manufacturers. We excluded manufacturers and models if there were less than 50 patients per each class of both manufacturer and model. We converted radiographic machine manufacturer' and model's names into unique numbers for confidentiality.

Radiographic technique

Radiographs of the knees were taken in one or more of the anterior-posterior (AP), posterior-anterior (PA) flexed, patellar, internal oblique, external oblique, and lateral projections. Ankle radiographs were taken in the AP, lateral, or mortise projections; while leg radiographs were taken in the AP and lateral projections. Foot radiographs were taken in the AP, lateral, and oblique projections. The radiographs originate from different sites (knees, legs, ankles, and feet) and were obtained in different projections with varying angles, so it is difficult to match the contents of images to the origin of the radiographic machines. Most patients had at least three images as part of their radiology study.

Image pre-processing

Radiographs were downloaded from the Visage Picture Archiving and Communications System (PACS) in Digital Imaging and Communications in Medicine (DICOM) format. These images were then converted to the Joint Photographs Expert Group (JPEG) format and resized to 256x256. The image pixel intensities were normalized to the range 0 to 255. We generate a batch of images by adding relevant weights to each instance during training process to deal with class imbalance between manufacturers and models. The weights are proportional to the frequency of the unique labels in the dataset. We apply the weights when we select instances for batches. At the end, we have a batch of data with a suitable number of images per class in every batch.

Experiment:

We used the deep learning model to predict the (1) radiographic machine manufacturer and (2) the radiographic machine model. We used the following deep learning architecture for both experiments.

Architecture

We use deep learning for radiographs source identification (Deep-RSI) to extract features on-line to generate a model on the train dataset in order to predict the radiographic machine source (manufacturer and model) in the test dataset. Figure 1 shows the order of the process: input, deep learning and output.

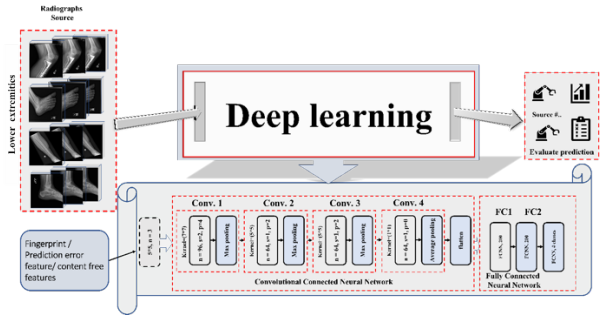


Figure 1: Content-free deep learning architecture for radiographs manufacturer and model Identification

Fingerprint feature extraction

Normally, digital images intensities are derived from three channels: red, green, and blue (RGB). The deep learning algorithm uses all the image content pixels' value which are called image intensities. These pixel intensities serve as inputs to the deep learning model. We convert gray-scale radiographs to single-channel images. We add a layer at the beginning of CNN layers to extract low-level features from content-free pixels using a kernel filter with a size of (5*5) (Figure 1). This layer uses 3 filters of the same size to perform like three-channel images which is also referred to as fingerprint, residual images, prediction error feature / content free features. This layer keeps only pixels (noise) that are sparsely distributed in the images. Majority of pixels in the images are black (i.e., the intensity values are close to zero). The extracted information from the content-free pixels is used as the input for the subsequent analyses. We use the CNN to extract 6272 high-level features from the low-level features.

CNN

We use four convolutional layers within the CNN. The first layer uses 96 filters with a filter (kernel) size of (7*7) and operates on stride=2 while taking padding (p =4) into account (Figure 1). Padding (p=1) is a zero-value pixel added to the periphery of the input image. When p is equal to four, it means that we add four pixels with normalized intensity of 0 to the periphery of the input image. Pixels at the edge of the images can be erroneously ignored during the feature extraction. Padding helps the CNN feature extractor to evaluate all pixels, including pixels located at the border/edge of images.

We compute all the filters to generate the initial high-level features in the first layer. We apply a max-pooling function with a configuration of (size = (2*2) and stride =2) to shrink the dimension of generated high-level features within each channel by half while keeping the number of channels (96).

The second layer uses the output of the first layer (96 channels) to apply 64 filters with stride =1, after applying padding(p=2). We use a filter size of (5*5) to generate 64 channels with the same input size. We apply the max-pooling with size = (2*2) and stride=2 to reduce the dimension of each channel by half.

In the third layer, we repeat the identical configuration of the second layer (the same filter count and size, stride=1 and padding=2). We also apply the max-pooling function with size = (2*2) and stride =2 to decrease the dimension of each channel by half in this layer as well.

In the final layer, we utilize 64 filters with a size of (1*1) with stride =1 and padding(p=0), which means that we do not add padding in this layer. We apply an average-pooling with size = (2*2) and stride = 2. Finally, we flatten the output of the fourth layer to prepare it for training process by FCNN.

FCNN

We leverage the FCNN as our main classifier with three layers (Figure 1). The first FCNN layer computes 6272 features using 200 nodes to generate weights and uses a rectified linear unit (ReLU) which is an activation function to filter input wights. If the input is positive, ReLU outputs it directly; otherwise, it outputs zero. The layer passes filtered weights to the second layer with 200 nodes with ReLU. The third layer employs a softmax function coupled with a value equal to the number of classes to build a probility of predicting output class. FCNN selects the class prediction with the greatest probability.

Software

We conduct this research in Python=3.10.8 using deep learning libraries: Pytorch=1.13.1, Pytorch-cuda=11.7 and Torch=1.13.1. We evaluate the hyper-parameters with different values to obtain the optimal validation learning rate values with 0.1, 0.01, 0.001, 0.0001. We choose the Adam optimizer and set a learning rate equal to 0.0001, epochs equal to 50, and batch size equal to 20.

Hardware

We run this research study on a system with the following configurations: GPU: NVIDA Quadro RTX 4000 (DRAM: 8GB), CPU: Intel® Xeon® W-2123, 3.60 GHz, and random-access memory (RAM):64 GB.

Model performance statistics

In this research, we use five key metrics to evaluate the Deep-RSI classification performance in both experiments: manufacturer identification and model identification. Statistics were calculated using the scikit-learn Python library.

1) Accuracy: This metric measures the overall performance of classification performance independent of classes using formula 1.

$$\text{Accuracy} = \frac{TP+TN}{P+N} \quad \text{(Formula 1)}$$

Where TP stands for true positive, TN is true negative, P stands for the number of positive samples and N stands for the number of negative samples.

2) Area under curve (AUC): We use this measure to calculate the ratio of false positive rate (FPR) against true positive rate (TPR). AUC also measures the whole area underneath the entire receiver operating characteristic (ROC) curve from (0,0) to (1,1) using formula 2. ROC plots a chart using the FPR (on x axis) and TPR (on y axis). FPR is the ratio of false positive over the sum of false positive and true negative.

$$\text{AUC} = \int_0^1 ROC \, dx \quad \text{(Formula 2)}$$

3) Specificity: This measure is also known as the true negative rate (TNR) where we look for negative samples which are correctly classified using formula 3.

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{TN}{N} \quad \text{(Formula 3)}$$

4) Sensitivity: This is the true positive rate (TPR) in which we measure the rate of correctly classified positive samples according to formula 4. Sensitivity is also known as recall.

$$\text{Recall} = \text{Sensitivity} = \frac{TP}{TP+FN} = \frac{TP}{P} \quad \text{(Formula 4)}$$

where FN is the number of positive samples incorrectly predicted negatives, and P is the number of positive samples.

5) Positive Predictive value (PPV) or Precision: This measurement is the ratio of the true positives to the number of true positives and false positives.

$$\text{PPV} = \text{Precision} = \frac{TP}{TP+FP} \quad \text{(Formula 5)}$$

6) Negative Predictive Value (NPV): NPV is the ratio of the number of true negatives to the sum of the true negatives and false negatives

$$\text{NPV} = \frac{TN}{TN+FN} \quad \text{(Formula 6)}$$

3. Results

Experiment 1

In experiment 1, we assessed the proposed method on radiographs manufacture classification for each site (Table 1).

Table 1: Experiment 1: Radiographic machine manufacturer classification for radiographs of the lower extremities

Site	Total Patients	Total Images	Train/ Validation (patients)	Train/ Validation (images)	Test (patients)	Test (images)	No of classes (test)	Weighted average Accuracy (images)	Weighted average AUC (images)
Knees	1418	6119	1132	4849	286	1270	4	0.96	0.95
Legs	616	2177	494	1726	122	451	3	0.95	0.93
Ankles	1290	4598	1038	3692	252	906	5	0.97	0.95
Feet	1074	3986	859	3183	215	803	3	0.99	0.99

Table 2: Experiment 1: Detecting the radiographic machine manufacturer for radiographs of the knees, legs, ankles and feet

Body site							
Knees							
Manufacturer	No. of samples	Accuracy	AUC	Specificity	Sensitivity	PPV	NPV
Class 0	237	0.97	0.93	0.99	0.88	0.94	0.97
Class 1	202	0.99	0.98	0.99	0.96	0.95	0.99
Class 2	45	0.99	0.93	0.99	0.87	0.95	0.99
Class 3	786	0.96	0.96	0.93	0.98	0.96	0.97
							Legs
Class 4	332	0.96	0.95	0.95	0.96	0.98	0.90
Class 5	48	0.94	0.86	0.97	0.75	0.73	0.97
Class 6	71	0.92	0.87	0.95	0.80	0.74	0.96
							Ankles
Class 7	25	0.99	0.99	0.99	1.0	0.78	1.0
Class 8	74	0.99	0.95	1.0	0.90	1.0	0.99
Class 9	82	0.97	0.96	0.98	0.94	0.80	0.99
Class 10	208	0.96	0.94	0.98	0.89	0.92	0.97
Class 11	517	0.99	0.99	0.99	0.98	0.99	0.98
							Feet
Class 12	199	0.99	0.99	0.99	0.98	0.98	0.99
Class 13	211	0.99	0.99	0.99	0.99	0.97	0.99
Class 14	393	0.99	0.99	0.99	0.98	0.99	0.98

Knee radiograph manufacturer analysis: Table 2 shows the results of knee radiographs' analysis for the four imbalanced classes. The proposed method obtained the highest *accuracy* with values of 0.99 and 0.99 for *class 1* and *class 2*, respectively. *Class 1* is the most easily identified manufacturer for classification of the radiographic machine used to obtain knee radiographs. The proposed method obtained the highest *AUC* for class 1 (AUC=0.98).

Leg radiograph manufacturer analysis: Class 4 is the most easily identified manufacturer for radiographs of the legs and our model had *accuracy, AUC, sensitivity* and *PPV* of 0.96, 0.95, 0.95, 0.96 and 0.98, respectively to detect radiographs created by Class 4.

Ankle radiograph manufacturer analysis: Ankles radiographs were all taken using 1 of 5 radiographic machine manufacturers. The proposed method had a high accuracy of 0.99 and AUC > 0.94 for identifying which radiographic machine was used to obtain ankle radiographs.

Foot radiograph manufacturer analysis: Again, our proposed method had > 0.99 *accuracy* and *AUC* to correctly predict which radiographic machine was used to obtain a patient's foot radiographs.

Experiment 2

In this experiment, we try to identify the model of the radiographic machine used to create radiographs of the knees, legs, ankles, and feet. Each dataset has a different number of classes: Knees and feet have 5 classes, legs have 3 classes, and ankles have 4 classes. We assessed model datasets and obtained results of average classification *accuracies* and *AUCs* for determining the model used for creating the knee, leg, ankle, and foot radiographs which are shown in Table 3. Table 3 redemonstrates that we can predict the radiographic machine model used to generate leg radiographs with accuracy and AUC of 0.99 and 0.99, respectively. Despite the imbalance of our dataset, both accuracy and AUC are high. This verifies that the proposed method produces a model that is not biased toward the dominant class.

Table 3: Experiment 2: Radiographic machine model classification for radiographs of the lower extremities

Site	Total Patients	Total Images	Train/ Validation (patients)	Train/ Validation (images)	Test (patients)	Test (images)	No of classes (test)	Average Accuracy (images)	Average AUC (images)
Knees	1326	5431	1070	4579	256	852	5	0.97	0.94
Legs	546	1888	437	1518	109	370	3	0.99	0.99
Ankles	1086	3849	952	3429	134	420	4	0.98	0.98
Feet	1074	3983	957	3132	117	851	5	0.97	0.94

Classification results for radiographics machine models across four sites are shown in Table 4. The experimental findings show that the model has high accuracy and AUC to identify the radiographic machine model used to create radiographs with values of 0.99 and 0.99 respectively.

Table 4: Detecting the radiographic machine model for radiographs of the knees, legs, ankles, and feet

Body site							
Knees							
Classes	No. of images	Accuracy	AUC	Specificity	Sensitivity	PPV	NPV
Class 0	75	0.95	0.83	0.98	0.68	0.78	0.97
Class 1	144	0.98	0.96	0.99	0.93	0.98	0.98
Class 2	26	0.99	0.90	0.99	0.81	0.84	0.99
Class 3	534	0.97	0.96	0.94	0.98	0.96	0.97
Class 4	73	0.95	0.86	0.96	0.75	0.67	0.98
Legs							
Class 5	70	0.99	0.98	0.99	0.97	0.96	0.99
Class 6	43	0.98	0.94	0.99	0.88	0.93	0.98
Class 7	257	0.99	0.98	0.97	0.99	0.99	0.98
Ankles							
Class 8	42	0.98	0.97	0.98	0.95	0.85	0.99
Class 9	26	0.99	0.99	0.99	1.0	0.96	1.0
Class 10	228	0.99	0.99	1.0	0.99	1.0	0.99
Class 11	124	0.97	0.96	0.99	0.93	0.97	0.97
Feet							
Class 12	105	0.93	0.84	0.96	0.72	0.72	0.96
Class 13	36	0.97	0.86	0.98	0.75	0.63	0.99
Class 14	174	0.97	0.95	0.98	0.91	0.93	0.98
Class 15	449	0.97	0.97	0.98	0.96	0.98	0.96
Class 16	87	0.94	0.85	0.96	0.75	0.70	0.97

Our model consistently could predict the model of the radiographic machine used to obtain radiographs of the knee, leg, ankles, and feet (Table 4).

4. Conclusion

We created a deep learning model to detect the make and model of the radiographic machine used to take radiographs with high accuracy and high AUC. Our model could reliably do this using 4 different datasets, with no change in accuracy or AUC between the training/validation dataset and the test datasets. We used the content- free pixel information to detect the radiographic machine source for radiographs of the different sites

in the lower extremities. This is a novel extension of previously published method based on analysis of content-free pixel information from digital cameras, applied to digital radiographs.

Radiograph source identification has scientific, medical, and legal implications. First, scientists may be able to spot scientific fraud, such as cut-and-paste radiographs from inconsistent sources. Second, if a healthcare system is hacked and fake radiographs are placed in a patient's medical record, then the patient's diagnosis and treatment may be altered. Third, if there is a medical malpractice case and the radiograph with a critical finding to the case is replaced by a fake radiograph, then justice may not be served, and the patient may not be compensated for the medical error. We developed a deep learning-based approach to make use of content-free pixel information as the first step to detect forgery via source identification solution.

There are series of research studies in which scientists used machine learning to train a model for source identification using extracted imaging features. Initially, researchers extracted a list of features followed by using support vector machine (SVM), artificial neural network (ANN) and random forest (RF) (Zhang et al, 2020). In (You et al, 2020) researchers looked for the sources of light using statistical fluctuations and correlation properties of features followed by using ANN and naïve Bayes (NB) classifiers for training and evaluating the trained model for sake of the source identification problem. Machine learning approaches were not efficient because of the manual feature extraction process ((You et al, 2020). There was a need for automatic feature extraction and learning strategies to improve efficiency. Deep learning is an effective way to extract features automatically and train the model based on these features simultaneously. Researchers used the trained model for source classification problem to classify test datasets (Freire-Obregon et al, 2019). Deep learning is an advanced learning technology in which we use a convolutional neural network (CNN) and classifier, a fully connected neural network (FCNN). The CNN extracts various numbers of features using different convolutional layers. Every layer shares a kernel filter, with an activation function, stride, and padding. The FCNN uses the collected features to create a model with a softmax function in the last layer for source prediction. Source prediction is an important first step in image authentication.

Researchers divided the image authentication approaches into two categories: active and passive authentication (Al-Qershi & Khoo, 2013). In active authentication, the goal is to identify an actively hidden object / actively placed watermark or digital signature in the images. The images are altered to confirm the originality of images by embedding a watermark image or using a signature (Al-Qershi & Khoo, 2013). These actively placed object is known and can be detected by the forensic expert. Passive authentication is different from active authentication because there is no actively placed object embedded in the images. Passive authentication can be used to identify an image where only a region within the image is altered. Identifying these fabricated regions is also challenging and requires region-based steganalysis approaches (Ghareh Mohammadi & Sajedi, 2017). We focus on the content-free pixels which are sometimes considered noise. Researchers have shown that the noise pixels contain information content that could be used to determine the device used to create an image (Ghareh Mohammadi & Sebro, 2022). In general, all types of medical/non-medical cameras leave identifiable noise while creating images/radiographs due to different capturing process and technologies. We exploit intrinsic noise in the radiographs in such a way that we can predict the source of the radiographs.

Researchers developed several approaches for extracting information from the noise in the digital images. This noise is referred to as content-free pixel information, and these are the primary pixels essential for solving this problem. Originally, researchers extracted the noise pixel information directly from images to identify the source of images. They utilized Photo Response Non Uniformity (PRNU) (McCleary, 2009), (Rosenfeld & Sencar, 2009) and Sensor Pattern Noise (SPN) (Lukas et al, 2006) to extract relevant information from images. Peak to Correlation Energy (PCE) ratio was used to assess the extracted noise information for source identification (Goljan et al, 2009), (Kang et al, 2011). Researchers even studied whether the noise varied between photographs that were obtained by a professional or amateur to determine if the noise was intrinsic to the device or the user (Rosenfeld & Sencar, 2009).

Prior studies have shown that there is a way to relate digital images/videos to the device used to create these digital images/videos (Al-Qershi & Khoo, 2013). However, there is little known about source identification of medical images. Advanced AI techniques, particularly deep learning-based approaches, were first used for MRI source identification (Fang et al, 2020), which is the first forensic research study of this type in the medical field. Ghareh mohammadi and Sebro in (Ghareh Mohammadi & Sebro, 2022) expanded this approach for radiograph source identification. They conducted a Deep-RSI approach to identify the manufacturer and model of the upper extremities' radiographs (Ghareh Mohammadi & Sebro, 2022). In this paper, we used a modified Deep-RSI approach for source identification of lower extremities radiographs.

The main limitation in this work is that there were a limited number of manufacturers and models used to generate the radiographs. This is the typical number of devices at a large tertiary care healthcare institution. However, in future we will investigate the performance of our model with larger number of manufacturers and models. Radiographs were taken from different angles, which may have influenced the performance of the model. If we have enough radiographs from all different angles, we might have higher performance. Finally, the data all come from a single tertiary care academic center, and the devices investigated may not generalize to all radiographic machines.

In summary, we use a deep learning model to investigate lower extremity radiographs to accurately predict the make and model of the machine that generated the radiograph.

References

Al-Qershi, O. M. & Khoo, B. E. (2013) Passive detection of copy-move forgery in digital images: State-of-the-art. *Forensic science international*, 231(1-3), 284-295.

Fang, S., Sebro, R. A. & Stamm, M. C. (2020) A deep learning approach to MRI scanner manufacturer and model identification. *Electronic Imaging*, 217-1 -- 217-6.

Freire-Obregon, et al (2019) Deep learning for source camera identification on mobile devices. *Pattern Recognition Letters*, 126, 86-91.

Ghareh Mohammadi, F. & Sajedi, H. (2017) Region based image steganalysis using artificial bee colony. *Journal of Visual Communication and Image Representation*, 44, 214-226.

Ghareh Mohammadi, F. & Sebro, R. (2022) Deep-RSI: Deep learning for Radiographs Source Identification *2022 international conference on computational science and computational intelligence (csci)*, Pressing.

Goljan, M., Fridrich, J. & Filler, T. (2009) Large scale test of sensor fingerprint camera identification, *Media forensics and security*. SPIE.

Goodfellow, I., et al. (2020) Generative adversarial networks. *Communications of the ACM*, 63(11), 139-144.

Jaiswal, A. K. & Srivastava, R. (2022) Detection of copy-move forgery in digital image using multi-scale, multi-stage deep learning model. *Neural Processing Letters*, 54(1), 75-100.

Kang, X., et al. (2011) Enhancing source camera identification performance with a camera reference phase sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 7(2), 393-402.

Lukas, J., Fridrich, J. & Goljan, M. (2006) Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2), 205-214.

McCleary, B. (2009) *Digital imaging system testing and design using physical sensor characteristics*University of Southern California.

Rosenfeld, K. & Sencar, H. T. (2009) A study of the robustness of PRNU-based camera identification, *Media Forensics and Security*. SPIE.

You, C., et al (2020) Identification of light sources using machine learning. *Applied Physics Reviews*, 7(2), 021404.

Zhang, H., et al (2020) Machine learning-based source identification and spatial prediction of heavy metals in soil in a rapid urbanization area, eastern China. *Journal of Cleaner Production*, 273, 122858.

Cyberspace Geography and Cyber Terrain: Challenges in Producing a Universal map of Cyberspace

Alexander Grandin
Finnish Defence Research Agency, Riihimäki, Finland

alexander.grandin@mil.fi

Abstract: Much in the same way that cyber has become the fifth military domain, cyberspace has also brought forth the research area of Cyberspace Geography. The challenge of producing a universal map of cyberspace however still exists. Cybersecurity specialists, military personnel and researchers still begin with a blank sheet on which the wanted elements of cyberspace are arranged before solving their actual problem. The abundance of elements in cyberspace requires a careful selection of factors to include in one's map, depending on how it will be used. However, a complex and ever-changing environment such as cyberspace could make use of a generally acknowledged starting point, facilitating this work. In previous research cyberspace has been described as a combination of the physical world, the social world and the information world. The multidisciplinary research in Cyberspace Geography has developed models for mapping and displaying cyberspace. This is often done by creating topological maps, much like the map of the New York subway system. Military cybersecurity researchers have through the concept of Cyber Terrain presented similar models of cyberspace for military operations. Research has also been produced on the techniques and methods for mapping cyberspace as well as the different presentations of the mapped information. Graph theory has for instance been used as a mathematical model of cyberspace. It is nonetheless unclear if there is some degree of universality in the elements that the different research presents. Which are e.g. the similar features between the cyberspace maps that are used for military operations, that describe the cyber environment of a country or between the elements used for modelling a cybersecurity system? This paper aims to present a solution to this challenge by systematically reviewing the research on Cyberspace Geography and Cyber Terrain using thematic analysis. The different elements of the maps of cyberspace are reviewed. The research will answer if a universal map, that can be used as a starting point for solving multiple challenges in cyberspace, can at present be prepared.

Keywords: Cyber Defence, Cyber Terrain, Cyberspace Geography, Cyber Territory, Cyberspace Mapping

1. Introduction to Cyber Terrain and Cyberspace Geography

According to Google Scholar[1], Cyberspace Geography has yearly appeared in publications since 1993 and Cyber Terrain since 1998. Interestingly, Cyberspace Geography experienced a small renaissance around the year 2019, with nearly as many search hits as under its previous highlight around 2000. This seems to be a result of for instance Chinese scholars using the term frequently since 2017. The term has not found an especially large audience, since only about 196 search hits appear altogether on Google Scholar compared to Cyber Terrain that produces 520 search hits.

Cyber Terrain however was more widely used after 2009, with most search hits between 2017 and 2018, as can be seen in figure 1. If Cyberspace Geography was at least in the beginning associated with geographers, Cyber Terrain research seem to be associated with the term cyberwar, cyberoperations and, at least in part, with military planning in western countries.

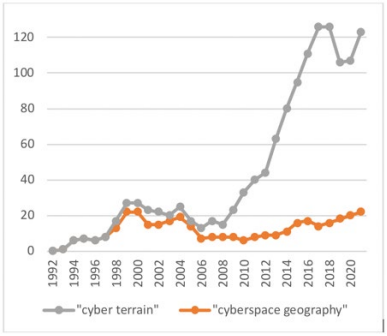


Figure 1: Search hits on Google Scholar, showing "cyber terrain" in gray and "cyberspace geography" in orange since 1992

[1] *The limitations of using Google Scholar were identified, but were outweighed by the possibility to gain an international and interdisciplinary approach to the chosen search terminology. For instance, EBSCO does not provide some of the Chinese research on cyberspace geography, that has been published on Springer. The point of view was thus chosen to be as broad as possible, even though it might have lacked in depth.*

Mills (2012) describes in his article the evolution of the traditional warfare concept of key terrain, derived from Carl von Clausewitz, to the concept of key cyber terrain. Mills argues that key cyber terrain has traditional components as well as contemporary nodes. For instance, a data centre has a physical dimension as well as a distinct cyber dimension. Key cyber terrain is a combination of traditional concepts of warfare adopted to cybersecurity, in the same manner that cyberspace geography is a bridge between traditional geography and cyberspace. The question is if these two traditional ways of looking at the physical realm also portray cyberspace in the same manner, and if that could form the basis for a multidiscipline map for cyberspace.

The term Cyberspace Geography has in recent years been connected to the term Cyberspace Surveying and Mapping (CSM) by researchers like Xu (2019) and Kou, Ni & Du (2022). Xu (2019) also connects the concepts of network mapping, cyberspace mapping and cyberspace cartography to the subject and treats these as identical. The concept of the "ternary world" has also been proposed at least as early as 2011 by Xu & Li. This concept of a world consisting of the physical world, the social-human world and the information world is further developed by Lü, Yuan & Yu (2021). The division into three spatialities or levels is conceptually quite near the US Joint Chiefs of Staff (2018) definition of the three layers of cyberspace, which are the physical network layer, the logical network layer and the cyber-persona layer.

Ferreira & Vale (2021) describes how cyberspace has been explored by geographers, and notes that the research has most notably used the terms "geographies of internet", "cyber geography", "geography of cyberspace" and "virtual geography". The recent revival of Cyberspace Geography, especially in China, seems to be well described by Kou, Ni & Du (2022), stating "*Studying cyberspace surveying and mapping technology and comprehensively mastering cyberspace characteristics and resource distribution have very important theoretical significance and practical value for ensuring national security.”*.

2. Previous research and realization

The systematic literature review was conducted using the keywords "Cyberspace Geography" and "Cyber Terrain" to find relevant research on the subjects. The primary source of information was Google Scholar. Articles and documents dating from 2002 were collected, which resulted in 10 sources on cyberspace geography and 24 on cyber terrain. The sources were thematically analysed in order to define how cyberspace was constructed by the authors. 5 sources on cyber terrain were deemed irrelevant and 8 sources did not provide information that defined what constitutes cyber terrain. On cyberspace geography, one source proved irrelevant and 3 sources did not provide required information. This left 17 sources for the thematic analysis. The sources were primarily research articles or conference papers, but also reports and theses (6 sources) and government publications (2 sources). The thematic analysis was done by listing the definitions and viewpoints on the subject and comparing these to find reoccurring elements and themes. Some, albeit few, of the analysed articles dealt directly with the task of defining (key) cyber terrain.

The concept of "key cyber terrain" has previously been assessed by Huntley (2016). Huntley (2016) notes that "*The concept of "key terrain" applied in cyberspace is necessarily metaphorical.*" and continues by stating that this simple fact is rarely noted. He reviews 11 studies or doctrinal publications, primarily by the U.S. Department of Defence or by actors tied to them. He concludes that key cyber terrain is primarily linked to the US Department of Defence (DoD) definitions of cyberspace and (physical) key terrain, and thus also to the physical aspects of cyberspace, such as hardware and infrastructure. He continues by noting that there seems to be a practical need to try to take advantage of existing military planning concepts, such as key terrain, and translate these into cyberspace, even though this might result in loosing valuable virtual aspects of cyberspace, such as cyber-persona and virtual spaces. Huntley also identifies that in the studies he examined, the concept of key cyber terrain varies on different operational levels, and concludes by questioning the usefulness of the concept and even the possibility to define the terrain in a useful way.

Martin Dodge and Rob Kitchin (2001) published the first book devoted to cyberspace geography, *Mapping Cyberspace*. The book provides an analysis of cyberspace from a geographer's point of view, but draws on research in e.g. cartography, sociology and information visualisation. The book and the research around cyberspace geography seem to be especially supported by the discussion around the year 2000 if geography was becoming obsolete, as ICT's eliminates the time and distance between spaces and places.

Boos (2017 pages 13-38) discussed in his book the term cyberspace by reviewing academic literature on human geography, sociology and cultural anthropology. The review is however strictly defined, and only focus on concepts of on- and offline life. He however states that since the virtual world of cyberspace is connected to the real world, there must also be different cyberspaces for different connections between the on- and offline world.

These also change over time as the topologies of information change. He also cites research that point out that cyberspace is mainly virtual and socio-cultural. This points out the difference between the reality of how cyberspace is conceived by users (mainly as a virtual social construction) and the cybersecurity community, that focus on how it is constructed, which highlights the physical and logical foundation of cyberspace.

3. Results from the thematic analysis

The analysis shows that cyber terrain remains difficult to define. Guion & Reith (2017) and Price et al. (2017) describe how key cyber terrain can be evaluated by subject matter experts. The lack of definitions in cyberspace has been discussed by Applegate, Carpenter & West (2017), identifying research that extensively list key cyber terrain, thus confusing it with the elements of cyberspace. They point out that previous efforts have three main flaws. First, they state that instead of *what* cyber terrain consists of, the research should focus on how *key* terrain can be identified. Secondly, the concept is not linked closely enough to military planning and thirdly, that critical assets and cyber key terrain are confused with each other. Cyber terrain and key cyber terrain are in multiple studies (e.g. Bodeau, Graubart & Heinbockel (2013), Raymond et al. (2014) and Jabbour & Muccio (2011)) described using the US Joint Chiefs of Staff publication 2-01.3 (JP 2) (2009) definition, which is "*Any locality, or area, the seizure or retention of which affords a marked advantage to either combatant*".

Price et al. (2017), Applegate, Carpenter & West (2017) and Franz (2012) makes the distinction between cyber terrain on the tactical, operational and strategic levels. It is unsurprising that cyber terrain should be redefined at different levels – a module of a software may be critical for a system, that in turn is critical for a network etc. Another emblematic feature of cyber terrain is the temporal linkage to a specific (military) mission, as presented by Bertoli & Raio (2018) and Franz (2012). Several researchers, as Guion & Reith (2017), Price et al. (2017) and Jackson (2016) highlight the role of subject matter experts in defining key cyber terrain, but offer little advice on how this should be done.

Key cyber terrain is e.g. by Williams (2014) exemplified by a system (e.g. the Ballistic Missile Defence System (BMDS)). Bertoli & Raio (2018) describe the difficulty of distinguishing between cyber terrain and cyberspace in their article, and find that cyber terrain usually equals the features that build up cyberspace. They however note, that compared to the physical world, where terrain is defined, cyber terrain is not defined as a finite subset of the cyberspace domain, but comprise of all features of the cyberspace domain. They also argue that the concept of key terrain is difficult to translate to cyberspace. They mention that if the physical representation of a system such as BMDS does not constitute key terrain, then how could the virtual part of the system be key cyber terrain

Jakobson (2011) defines cyber terrain by a structure consisting of the hardware, software and service sub-terrains. The hardware sub-terrain is constructed of connected network infrastructure and devices, including information such as location and connectivity for these. The software sub-terrain consists of e.g. operating system and applications. The software components could have additional information such as vendor, release and known vulnerabilities. The service sub-terrain consists of the services and their intra-dependencies, including databases, e-mail, universal time, etc. The definition that Jakobson presents evades to shine light on what the difference between cyber terrain and cyberspace is, as his definition is similar to that of cyberspace. He however does not identify cyber persona as a factor of cyber terrain.

Mills (2012) define key cyber terrain primarily as consisting of physical objects, such as data centres, undersea information cables and internet service providers (ISP:s) but also mention BIOS (Basic Input-Output System), cyber workforce, supply chains, international standards bodies and innovation. He states that the Clausewitzean key terrain concept holds for cyber key terrain, because of its physical manifestations. He mentions "*geographic duplicity of data stores*" but still pinpoint data centres as key cyber terrain. Interestingly, he highlights the importance of cyber workforce, international standards bodies and innovations. Innovations, workforce and international standards bodies hardly change the outcome of any specific mission or operation in cyberspace. Mills seem to highlight important features and assets on a strategic level, that may impact the bigger picture of a nation-state's capabilities in cyberspace. Franz (2012) presents similar examples of key cyber terrain as Mills (2012) and names optic cables, satellite communications, subnets, databases with usernames and passwords and technicians in his work.

Raymond et al. (2014) defines cyber terrain through the definition of cyberspace developed by Raymond et al. (2013). By their definition, cyber terrain consists of "*the systems, devices, protocols, data, software, processes, cyber personas, and other networked entities that comprise, supervise, and control cyberspace*." They define key cyber terrain similarly to JP 2. By this definition, all of cyberspace can at some point be defined as key terrain, depending on the mission and situation. A very similar description of cyber terrain is offered by Lemay, Knight

& Fernandez (2014). They use direct analogies to the physical terrain in their article and define cyber terrain as based on "*material and logical infrastructure*". In the same way they define key terrain as that which is of essence for the task at hand, but offer little definition of cyber terrain in general.

Phister (2010) also defines cyber terrain through an analogy to physical terrain. He states that cyber terrain "*…is primarily hardware (i.e., switches, routers) and software (i.e., operating systems, protocols)"*. He also presents the analogy of cyber paths, which he describes as communication links connecting the various entities. He however does not highlight cyber persona or actors in cyberspace in his definition.

Moore (2004) provides seven categories through which a network environment can be described. These are (1) network classification (which networks are used in the operation), (2) architecture types (VPN:s, ethernet, wireless networks etc.), (3) software applications, (4) operating systems, (5) existing vulnerabilities, (6) types of information (stored or processed in the identified networks) and (7) the network activity baseline.

Since the publication of the earlier mentioned report by Huntley (2016), a new version of the US Joint Chiefs of Staff publication 3-12 *Cyberspace Operations* (JP 3-12) was released (2018) which broadens the definition of key cyber terrain so that it expands into the virtual realm of cyberspace, stating "*An additional characteristic of terrain in cyberspace is that these localities have a virtual component, identified in the logical network layer or even the cyber-persona layer*". La Pierre (2020) further broadens the definition by including elements of the Electromagnetic Spectrum (EMS) and needed power sources. He also deepens the definition of the cyber persona layer, naming the focus of this layer as "*virtual actors and how they relate to real individuals or organisations, as well as how they interact with each another, and with the network."*, citing the Danish doctrine for military cyberspace operations (2019) and Craig Jones (2015).

Youn et al. (2021) operationalize the three-layer definition of cyberspace into elements. The physical network layer includes the geographical location for network devices as well as hardware, software and infrastructure. Critical information storage locations are also included. The logical network layer is the virtual space of cyberspace. They expand the social and cyber-persona elements into the logical layer, by naming websites of influence and social platforms as elements of it. The cyber-persona layer includes IP-addresses and personal information about users. They also include organizational elements and structures in this layer.

The US JP 3-12 publication defines cyberspace (2018) as consisting of three layers, which are the physical network layer, the logical network layer and the cyber-persona layer. The physical network layer consists of "*the IT devices and infrastructure in the physical domains that provide storage, transport, and processing of information within cyberspace, to include data repositories and the connections that transfer data between network components*." This includes the geographic location of devices, such as hardware and infrastructure. The logical network layer consists of the connected elements that are "*abstracted from the physical network*". They highlight that the elements of this level can "*can only be engaged with a cyberspace capability*". They thus conclude that information stored in the logical network cannot be affected by physical means, as previous studies suggest, due to the virtual nature of storage. This is naturally not always the case, as data or information can be stored in a single facility, but can be deemed a logical starting point. The cyber-persona layer is constructed by virtual identities that are derived from the logical network plane, such as e-mail addresses and IT user accounts. Their definition of cyberspace is "*A global domain within the information environment consisting of the interdependent networks of information technology infrastructures and resident data, including the Internet, telecommunications networks, computer systems, and embedded processors and controllers."*. This definition was according to Phister (2010) presented as early as 2008.

Different definitions of cyberspace are abundant. For instance, Gao et al. (2019) citing Boos (2017), defined cyberspace as consisting of five levels, which are the geographical layer, the physical network layer, the logical network layer, the cyber-persona layer and the persona layer. These layers consist according to Gao et al. (2019) of geospatial support elements, IT equipment, infrastructure, data, applications and network processes, user accounts and physical users. A similar definition was presented by Franz (2012), naming them the geographic layer, physical infrastructure, the information layer, cyber identity and people.

Raymond et al. (2013) defines cyberspace as consisting of five planes, which are the supervisory plane, the cyber persona plane, the logical plane, the physical plane and the geographic plane. The geographic plane consists of physical resource locations, such as infrastructure (including power etc.). The physical plane includes hardware and its details (serial number and attached devices) as well as communication protocols, and the other aspects of the ISO Open System Interconnection (OSI) models' layer 1. The other layers of the OSI model belongs to the logical plane. Users virtual identities lie above the logical plane and forms the cyber persona plane. The authors

have interestingly added a supervisory plane, that other models don't identify as an own plane. This plane consists of the persons and systems "*that provide the command and control necessary to start, stop, or redirect cyber weapons*". Jackson (2016) use the same cyberspace definition but uses it to define key cyber terrain. He however does not present the features in greater detail.

Ferreira & Vale (2021) approach cyberspace from a geographic point of view. They suggest, as does Boos (2017), that the notion of one cyberspace should shift to multiple cyberspatialities undergoing digital transformation. They present a flora of concepts that portray cyberspace such as the geographies of internet, cyber geography, geography of cyberspace and virtual geography. The authors also cite Graham (2013) and present the notion that cyberspace is a geographical metaphor, such as Huntley (2016) described cyber terrain. They also contend that cyberspace is influenced by the Cyber Divide, as does Boos (2016), by which different assemblages of data, people, devices, infrastructure and space creates, according to the authors, different cyberspatialities for different users.

Grant (2014) approach cyberspace from a combined military-geographical starting point. He presents a layered ontology meant to support military command & control processes. He includes the geographical layer, the physical layer, the information layer, the cognitive layer and the socio-organizational layer in his ontology. The geographical layer consists of physical locations and the paths between these. The physical layer consists of the physical objects of interest (e.g. devices and humans) and the links between these, such as cables and wireless transmission links. These have one or several locations on the geographical layer. Devices and links usually have a physical and virtual location. The information layer presents data and information, including the software processing it. The cognitive layer presents knowledge stored and relayed by humans. Knowledge and information may have a location based on the person possessing knowledge or the media containing the information. The authors separate knowledge into beliefs and operators. The socio-organizational layer presents the relationships between persons belonging to a group. Groups using ICT systems may have a physical and a virtual location.

Lü, Yuan & Yu (2021) further develops the concept of a ternary world. This concept divides the world into the physical, socio-human and the information world. The physical world are the traditional, physical domains of land, sea, air and space and the socio-human world the social relationships and civilizations that humans form. The information world links these to each other. The surveying and mapping of the information world has according to the authors not been developed. They propose that seven geographical elements and seven social-human elements should be used to form information geography.

Xu et al. (2019) use what they call the generally accepted three-domain model of cyberspace, consisting of the physical, logical and cognitive level. These levels include nodes and edges. The nodes and edges and entities are presented in table 1 below. Relationships and attributes are connected to the entities to form a description of cyberspace using graphs.

Table 1: Cyberspace features divided into three layers according to Xu et al.

	Physical Level	Logical Level	Cognitive Level
Nodes	network equipment, countries, cities	IP address, network applications, network services	identity, e-mail address, ID number
Edges	connection relationships between nodes	logical connections such as VPN, switching, peer-to-peer applications	interactions between nodes
Entities	devices and resources that form the network infrastructure	application services	media and people using physical level entities to participate on logical level services

Kou, Ni & Du (2022) approach cyberspace by defining targets for CSM. They classify targets into physical and virtual assets, with the goal to "*obtain comprehensive and complete information of various elements in cyberspace.*". The physical assets include firewalls and computers etc. but also intelligent wear and Internet of Things (IoT) devices. Of these assets, information, location and topological relationships are collected. Virtual assets are according to the authors mainly virtual human information, such as photos, e-mail and social media information. Zhao, Luo & Liu (2016) use the same division of physical and virtual assets when listing tools for CSM. They also include organizational structures and avatars as part of the virtual assets. According to them, the main objective of CSM of physical assets is defining a geographical location and of virtual assets to map social space.

4. Conclusions

The challenge of defining cyber terrain, cyberspace and the possible differences between them, remains. As emphasized by cyberspace geographers, cyberspace is heterogenic and changes constantly. The layers of cyberspace seem to have different inertia, and the geographic information usually changes more slowly than the virtual information. It is remarkable how little the temporal aspect of cyberspace, and how time affects the different levels or planes, is covered in research. The research on cyber terrain identifies the effect of time as part of its operational focus, but no clear description of these effects were found. The temporal dimension of cyberspace would require more research.

The definitions of cyber terrain and cyberspace agree on including physical devices, geographic location and the logical layer, but seems to differ on virtual representations. Though tree layers are often used to describe cyberspace, five layers is frequently used as well. The trend in cyber terrain research seems to be towards including cyber persona and other virtual aspects of cyberspace. Most researchers agree on presenting information and data to some extent, but for instance socio-organizational information or cyber persona information divides the research. The research on cyberspace geography, especially the research linked to the concept of the ternary world, seems to highlight cyber persona and socio-organizational aspects. Some expand cyber terrain into the electromagnetic spectrum or human knowledge spheres, without widespread support. It however seems clear that since cyberspace is a human construction, the definitions of it change depending on the portrayed need.

There seems to be a need to use the traditional military definitions and processes created to handle conflict in the Cartesian space also in cyberspace. This might seem practical for the military establishment, but some researchers question if valuable aspects of cyberspace are neglected this way. Especially the metaphor of cyber terrain is questioned – is there a clear need to define this analogy to physical terrain, or would it be more beneficial to use the concept of *key* cyber terrain as a metaphor for what in cyberspace is useful for the mission at hand. There is however call for a clear definition of which elements can be integrated into key cyber terrain and which cannot. It seems easier to define what cyberspace is, and more difficult to define what it at least is not. The tacit knowledge of subject matter experts seems to be essential as long as definitions and processes are developed.

Further research is needed to identify the tools and techniques for mapping cyberspace. Some previous research has been identified, and there are multiple identified tools available. It is not clear if all aspects and levels of cyberspace can efficiently be mapped. Furthermore, the collected information should be stored in an efficient way and presented when needed. These considerations would also benefit from further research.

References

Applegate, S. D.;Carpenter, C. L.;& West, D. C. (2017). Searching for digital hilltops, A Doctrinal Approach to Identifying Key Terrain in Cyberspace. *Joint Force Quarterly 84.1*, 18-23.

Bertoli, G.;& Raio, S. (2018). The elusive nature of cyber terrain. *Journal of Cyber Security and Information Systems*, 40-47.

Bodeau, D.;Graubart, R.;& Heinbockel, W. (2013). *Mapping the Cyber Terrain; Enabling Cyber Defensibility Claims and Hypotheses to Be Stated and Evaluated with Greater Rigor and Utility .* Bedford, MA: the MITRE corporation.

Boos, T. (2017). Geographies of Cyberspace: Internet, Community, Space, and Place. In *Inhabiting Cyberspace and Emerging Cyberplaces. Geographies of Media* (ss. 13-38). Palgrave Macmillan.

Dodge, M.;& Kitchin, R. (2003). *Mapping Cyberpsace.* London: Routledge.

Ferreira, D.;& Vale, M. (2021). From cyberspace to cyberspatialities? *Fennia 199(1)*, 113-117.

Franz III, G. J. (August 2012). *Effective Synchronization and Integration of Effects Through Cyberspace for the Joint Warfighter.* Baltimore: presentation at AFCEA TechNetLand Forces-East Conference in Baltimore, MD: Armed Forces Communications and Electronics Association.

Gao, C.;Guo, Q.;Jiang, D.;Wang, Z.;Fang, C.;& Hao, M. (2019). Theoretical basis and technical methods of cyberspace geography. *Journal of Geographical Sciences, volume 29*, 1949-1964.

Grant, T. (2014). On the Military Geography of Cyberspace. *Proceedings, 9th International Conference on Cyber Warfare & Security (ICCWS 2014)* (ss. 66-76). Purdue University, West Lafayette, IN, US: ICCWS.

Guion, J.;& Reith, M. (2017). Cyber terrain mission mapping: Tools and methodologies. *2017 International Conference on Cyber Conflict (CyCon U.S.)* (ss. 105-111). Washington D.C.: IEEE.

Huntley, W. L. (2016). *Cyber Key Terrain: A Conceptual Assessment.* U.S. Naval Postgraduate School.

Jabbour, K.;& Muccio, S. (2011). The Science of Mission Assurance. *Journal of Strategic Security, Volume IV Issue 2* , 61-74.

Jackson, N. (2016). *The Cyber War: Maintaining and Controlling the "Key Cyber Terrain" of the Cyberspace Domain .* Maxwell Air Force Base, Alabama: Air Command and Staff College, Air University.

Jakobson, G. (2011). Mission Cyber Security Situation Assessment Using Impact Dependency Graphs. *14th International Conference on Information Fusion* (ss. 1-8). Chicago, IL, US: IEEE.

Joint Chiefs of Staff. (2009). *Joint publication 2-01.3, Joint Intelligence Preparationof the Operational Environment.* Washington DC: Joint Chiefs of Staff.

Joint Chiefs of Staff. (2018). *Joint publication 3-12, Cyberspace Operations.* Washington DC: Joint Chiefs of Staff.

Kou, W.;Ni, L.;& Du, J. (2022). Research on Technical System for Cyberspace Surveying and Mapping. *Advances in Artificial Intelligence and Security. ICAIS 2022. Communications in Computer and Information Science, vol 1587.* (ss. 566-574). Springer.

La Pierre, H. (2020). *Intelligence Preparation Of The Operating Environment: Building Towards A Better Understanding Of Cyber Operations.* Toronto, Canada: Canadian Forces College, Collège des Forces Canadiennes.

Lemay, A.;Knght, S.;& Fernandez, J. (2014). Intelligence Preparation of the Cyber Environment (IPCE): Finding the High Ground in Cyberspace. *Journal of Information Warfare, Vol. 13, No. 3*, 46-56.

Lü, G.;Yuan, L.;& Yu, Z. (2021). Information geography: A new fulcrum of geographic ternary world. *Science China Earth Sciences, 65(2)*, 383-386.

Mills, J. R. (2012). The Key Terrain of Cyber. *Georgetown Journal of International Affairs,* , 99-107.

Moore, C. J. (2004). *Preparing the Virtual Battlefield for War: A Cyber Threat "Survival Kit" for Commanders.* Newport, RI, USA: Naval War College.

Phiter, P. W. (2010). Cyberspace: the Ultimate Complex Adaptive System. *the International C2 Journal vol. 4 num 2*, 1-30.

Price, P.;Leyba, N. A.;Gondree, M.;Staples, Z.;& Parker, T. (2017). Asset Criticality in Mission Reconfigurable Cyber Systems and its Contribution to. *Proceedings of the 50th Hawaii International Conference on System Sciences.* Hawaii.

Raymond, D.;Conti, G.;Cross, T.;& Fanelli, R. (2013). A Control Measure Framework to Limit Collateral Damage and Propagation of Cyber Weapons. *2013 5th International Conference on Cyber Conflict (CYCON 2013)* (ss. 1-16). Tallinn, Estonia: IEEE.

Raymond, D.;Conti, G.;Cross, T.;& Nowatkowski, M. (2014). Key Terrain in Cyberspace: Seeking the High Ground. *6th International Conference on Cyber Conflict (CyCon 2014)* (ss. 287-300). Tallinn: IEEE.

Williams, B. T. (2014). The Joint Force Commander's Guide to Cyberspace Operations. *Joint Force Quarterly 73*, 12-15.

Xu, R.;Zhang, Z.;Rao, Z.;Chen, J.;Li, M.;Liu, F.;& Pan, S. (2019). Cyberspace Surveying and Mapping: Hierarchical Model and Resource Formalization. *IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (ss. 68-72). IEEE.

Xu, Z.;& Li, G. (2011). Computing for the masses: Virtual extension. *Communications of the ACM*, 129-137.

Youn, J.;Oh, H.;Kang, J.;& Shin, D. (2011). Research on Cyber IPB Visualization Method based on BGP Archive Data for Cyber Situation Awareness. *KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS VOL. 15, NO. 2*, 749-766.

Zhao, F.;Luo, X.;& Liu, F. (2016). Research on cyberspace surveying and mapping technology. *Chinese Journal of Network and Information Security 9.2* , 1-11.

Detect, Deny, Degrade, Disrupt, Destroy, Deceive: Which is the Greatest in OCO?

Tim Grant
R-BAR, Benschop, The Netherlands

tim.grant.work@gmail.com

Abstract: In the cyber kill chain literature, possible courses of action are listed as *detect*, *deny*, *degrade*, *disrupt*, *destroy*, and *deceive* (a.k.a. "the 6Ds"). These verbs denote defensive action to be taken against an intruder. By comparison, military doctrine for cyberspace operations encompasses cyberspace exploitation and attack, as well as defence. The question arises whether the 6Ds are also applicable to offensive action, i.e. exploitation and attack, or whether additional action verbs are needed. Military doctrine is evolving towards all-domain operations, in which action in cyberspace is integrated with action in the physical domains of land, sea, air, and space. This prompts the question as to whether the 6Ds are also suited to action in a physical domain. A pilot study of actual military operations that integrated cyber and physical action suggests that deception, delay, and denial of organisational and cyber entities is suited to cyber action, while seizure, capture, and destruction of physical entities is suited to physical action. Preference among action verbs may indicate when it is best to engage targets using cyber or physical resources and which action is preferred. This paper identifies which action verbs are best suited to offensive cyber operations in the context of all-domain operations. The paper reviews related theory on cyberspace and the cyber kill chain. It identifies action verbs in US Department of Defense (DoD) doctrine on information and cyberspace operations, comparing them to those in the US DoD Dictionary of Military and Associated Terms. After discussing the findings, the paper draws conclusions and recommends further work.

Keywords: Action verbs, strategic effects, offensive cyber operations, computer network attack, all-domain operations.

1. Introduction

1.1 Motivation

The cyber kill chain approach (Croom, 2010) introduced a proactive strategy for defending computer-based systems from intrusion. By focusing on the threat element of risk (namely the attacker), rather than on the system's vulnerabilities, the defender can cut off the attack before it has achieved its goals. This makes it possible to defend against intrusive attacks in real time and, in particular, against Advanced Persistent Threats (APTs).

The approach centres on the sequence of phases that an attacker must pass through to achieve his/her objectives (Croom, 2010): *reconnaissance*, *weaponization*, *delivery*, *exploitation*, *installation*, *command and control*, and *actions on objectives*. This sequence is the attacker's *kill chain*. The defender can interrupt the kill chain in any phase by taking appropriate action. Hutchins, Cloppert & Amin (2011, p.117) categorize the defender's possible courses of action as *detect*, *deny*, *degrade*, *disrupt*, *destroy*, and *deceive* (a.k.a. "the 6Ds").

In this paper, we focus on the attacker's viewpoint. The question then arises as to whether the 6Ds completely describe an intruder's possible courses of action. Military doctrine assumes that some or all of the 6Ds apply to offensive cyber operations (OCO), covering both exploitation (a.k.a. intelligence gathering and analysis, manoeuvre, and enabling actions) and attack. For example, US cyberspace operations doctrine identifies *deny*, *degrade*, *disrupt*, *destroy*, and *manipulate* as actions that can be taken in cyberspace attack (US DoD, 2018, p.II-7). More widely, the US Department of Defense (DoD) Dictionary of Military and Associated Terms (US DoD, 2021, p.2) lists 32 strategic effects that denote all possible military actions. This begs the question as to how many of these verbs are applicable to the cyberspace domain and which are applicable only to the physical domains (land, sea, air, and space). The answer is important for targeting, especially in all-domain operations (ADO) where the right military asset and the right action must be chosen to engage a target in any of the five domains (Borne, 2019).

A pilot study (Grant & Kantola, 2021) of actual military operations that integrated cyber and physical ("kinetic") actions in an offensive operation suggests that cyber action is suited to deception, delay, and denial of organisational and cyber entities, such as military units, user accounts, and stored data. By contrast, kinetic action is suited to the seizure, capture, and destruction of physical entities, such as territory, buildings, and equipment. It is noteworthy that this pilot study already identified three verbs (*delay*, *seize*, and *capture*) not listed in the 6Ds. These suggestive results need to be checked against OCO-related doctrine.

1.2 Purpose, scope, and paper structure

The purpose of this paper is to identify which action verbs are best suited to OCO in the context of ADO, where OCO includes both computer network exploitation and computer network attack. It contributes to meeting Lin's (2009) and Denning & Denning's (2010) calls for an open discussion of OCO in the scientific literature. The focus is on OCO performed by professional military and law enforcement organisations at the operational and tactical levels. It assumes that the operation is aimed at penetration of the target system(s).

This paper contains six sections. After an introductory section, Section 2 reviews related theory on cyberspace and the cyber kill chain. Section 3 reviews the relevant US military doctrine. Section 4 maps the action verbs found in the doctrinal publications to the 32 verbs from the US DoD Dictionary of Military and Associated Terms. Section 5 discusses the findings, and Section 6 draws conclusions and recommends further work.

2. Related Theory

2.1 Cyberspace

The US DoD defines cyberspace as the part of the information environment consisting of the interdependent network of ICT infrastructures and data (US DoD, 2018). It includes the Internet, communications networks, computer systems, embedded processors and controllers, and the information residing in and passing through these networks and systems. Cyberspace operations are the employment of cyberspace capabilities aimed at achieving objectives in or through cyberspace, and may be defensive or offensive in nature.

In military doctrine, cyberspace is typically modelled as a set of distinct, yet interrelated layers. We adopt UK doctrine because this covers both physical and cyber domains, as needed in ADO. There are six layers (UK MoD, 2016): *social*, *people*, *persona*, *information*, *network*, and *real*, with the real layer being divided into *physical* and *geographical* aspects.

Table 1: Layers in physical and cyber domains (adapted from UK MoD, 2016).

Layer	Entities found in this layer	Domain
Social	Organisations, groups & teams; human-human interaction; processes & SOPs; organisational culture.	Physical
People	Individuals (users, developers, administrators, maintainers).	
Persona	Accounts (user, e-mail, etc).	Cyber
Information	Data, applications & protocols; domain names; logical connections between nodes.	
Network	Network nodes and links.	
Real (with 2 aspects): - Physical - Geographical	Infrastructure; buildings; devices; cables; wireless & optical communication links. Locations of infrastructure, devices, cables, & communication links, and of individuals, groups, teams & organisations.	Physical

Inter-entity relationships often cross layers and may combine physical and cyber domains. For example, accounts (*persona* layer) are owned by individuals (*people* layer). Manned weapon systems, such as armoured vehicles, ships, or aircraft, carry teams (*social* layer) of individuals (*people* layer), are constructed from physical devices (*real* layer, *physical* aspect), and can be found at a geographical location (*real* layer, *geographical* aspect). They can be fitted with Command & Control (C2) and communications systems consisting of hardware (*real* layer, *physical* aspect) and software plus data (*information* layer) connected in a network (*network* layer), with login accounts (*persona* layer) for the individuals. Moreover, in sensor and weapon control systems and in C2 systems, data (a cyber entity in the *information* layer) may represent or stand in for entities in the physical domain. For example, data in a C2 system may represent a real-world aircraft, with its identity and type, its geographical location, altitude, speed, and direction of travel, and whether it is hostile, friendly, or neutral.

2.2 Cyber kill chain

The term *kill chain* is a military concept describing the attack process (irrespective of which domain(s) it occurs in) as a chain of actions. If any link in the chain is broken, then the attack fails. Lockheed Martin applied the concept to defending computer-based systems in Croom (2010). The cyber kill chain is detailed in Table 2.

Table 2: The components of the cyber kill chain (Croom, 2010, p.54).

Phase	Description
Reconnaissance	Research, identification and selection of targets, often represented as crawling Internet Web sites looking for email addresses or information on specific technologies.
Weaponization	Coupling a remote access Trojan with an exploit into a deliverable payload, typically using an automated tool. Increasingly, data files such as Microsoft Office documents or Adobe PDF files serve as the weapon delivery device.
Delivery	Transmission of the weapon to the target. The three most prevalent delivery vectors for weaponized payloads are e-mail, Web sites, and USB removable media.
Exploitation	Triggering of the attacker's code. Most often, the weapon exploits an application or operating system vulnerability. It might simply exploit the user persuading him to open an executable attachment, or leverage a feature of the operating system that auto-executes code.
Installation	Installing a remote access Trojan or backdoor on the victimized system, allowing the attackers to affect all users of the system and to maintain persistence across system reboots.
C2	Accomplished most often with an outbound beacon to an Internet controller server, which establishes the C2 channel. This connection provides the manual "hands-on-the-keyboard" access that is required by most APT malware.
Actions on objectives	The final stage required for a successful intrusion. The most common objective is data exfiltration: collecting, encrypting, and stealing information from the compromised system. Attackers might also seek to violate data integrity or availability. Yet another objective might be to move laterally through the victim's IT environment, spawning new kill chains on subsequent targets.

Hutchins, Cloppert & Amin (2011) described in detail the Lockheed Martin model for the cyber kill chain and its application to cyber defence, particularly against campaigns of multiple attacks (as would occur with APTs). It introduced the defender's courses of action, listed as the verbs *detect*, *deny*, *disrupt*, *degrade*, *deceive*, and *destroy* (a.k.a. "the 6Ds"). Hutchins et al did not define the 6Ds.

Hutchins et al (2011) related the 6Ds to cyber kill chain phases in the form of a matrix; see Table 3. The abbreviations are explained follows (ibid., p.117): "*This matrix depicts in the exploitation phase, for example, that host intrusion detection systems (HIDS) can passively detect exploits, patching denies exploitation altogether, and data execution prevention (DEP) can disrupt the exploit once it initiates. Illustrating the spectrum of capabilities defenders can employ, the matrix includes traditional systems like network intrusion detection systems (NIDS) and firewall access control lists (ACL), system hardening best practices like audit logging, but also vigilant users who can detect suspicious activity.*"

Table 3: Courses of action matrix (Hutchins et al, 2011, Table 1, p.117).

Phase	Detect	Deny	Disrupt	Degrade	Deceive	Destroy
Reconnaissance	Web analytics	Firewall ACL				
Weaponization	NIDS	NIPS				
Delivery	Vigilant user	Proxy filter	In-line AV	Queuing		
Exploitation	HIDS	Patch	DEP			
Installation	HIDS	"chroot" jail	AV			
C2	NIDS	Firewall ACL	NIPS	Tarpit	DNS redirect	
Actions on objectives	Audit log			Quality of service	Honeypot	

There have been criticisms of the Lockheed Martin model of the cyber kill chain, but these criticisms are irrelevant to this paper.

3. Relevant doctrine

Doctrine is as important to (military) practitioners as theory is to scientists. Military doctrine expresses the military's institutional culture, the way it fights, and its relationship with the society and state that sustains it (Jackson, 2013).

We restrict our research to US DoD doctrine. The authoritative public-domain source for current doctrinal publications is the US Joint Chiefs of Staff Doctrine website (https://www.jcs.mil/Doctrine/). We found action verbs in doctrine on information operations (IO), cyberspace operations (CO), and countering threat networks (CTN). No US DoD doctrine has been published for ADO.

3.1 Information operations

Hutchins et al (2011, p.5) explicitly states that the 6Ds came from the 2006 version of Joint Publication (JP) 3-13 *Information Operations* (US DoD, 2006). This version not only includes the 6Ds but also five additional verbs: *exploit*, *influence*, *protect*, *restore*, and *respond*. Hutchins et al do not explain why they excluded the additional verbs, despite most of them being obviously applicable to CO. All 11 verbs are defined as shown in Table 4.

Table 4: Actions defined in JP 3-13 (US DoD, 2006), p.I-9 & I-10.

Action	Definition
Destroy	To damage a system or entity so badly that it cannot perform any function or be restored to a usable condition without being entirely rebuilt.
Disrupt	To break or interrupt the flow of information.
Degrade	To reduce the effectiveness or efficiency of adversary C2 or communication systems, and information collection efforts or means. Information operations can also degrade the morale of a unit, reduce the target's worth or value, or reduce the quality of adversary decisions and actions.
Deny	To present the adversary from accessing and using critical information, systems, and services.
Deceive	To cause a person to believe what is not true. Military deception seeks to mislead adversary decision makers by manipulating their perception of reality.
Exploit	To gain access to adversary C2 systems to collect information or to plant false or misleading information.
Influence	To cause others to behave in a manner favourable to own forces.
Protect	To take action to guard against espionage or capture of sensitive equipment and information.
Detect	To discover or discern the existence, presence, or fact of an intrusion into information systems.
Restore	To bring information and information systems back to their original state.
Respond	To react quickly to an adversary's or others' information operations attack or intrusion.

JP 3-13 was subsequently updated in 2012 and 2014, but no action verbs appear in these later versions. Instead, readers are referred to JP 3-12 for details on CO. This is because CO – then termed *computer network operations* – were included in IO in 2006. Around 2010, CO was split off from IO when US Cyber Command was established (Theohary, 2022).

More recently, the doctrinal emphasis has shifted to the generation and use of information in joint operations. In September 2022, JP 3-13 was replaced by JP 3-04 *Information in Joint Operations*, but this document is not available to the public (Pomerleau, 2022).

3.2 Cyberspace operations

The first version of JP 3-12 *Cyberspace Operations* was published in 2013, but was classified. An unclassified update was published in 2018, and this is the version used in this paper (US DoD, 2018). JP 3-12 has since been updated in January 2022, but this version is not available to the public.

JP 3-12 (US DoD, 2018) has four chapters. Chapter 1 is introductory, outlining the nature of cyberspace and setting the framework for using cyberspace forces and capabilities. Chapter 2 is most relevant to this paper as it describes the core activities in CO. Action verbs are mentioned only in the context of OCO. By contrast, Chapters 3 and 4 describe the US-specific CO organisation and procedures, and are therefore not relevant.

JP 3-12 identifies three main cyberspace missions: operating the US DoD's own ICT infrastructure ("blue" cyberspace), and defensive and offensive CO. OCO are missions intended to project power in and through foreign cyberspace. Foreign cyberspace may be "red", denoting cyberspace controlled by the adversary, or "grey", meaning neutral cyberspace that is neither blue nor red. OCO targets adversary cyberspace functions or creates first-order cyber effects that cascade into the physical domains. Offensive actions may rise to the level of the use of force, up to physical damage to or destruction of the adversary's systems.

There are two types of offensive cyber action: exploitation and attack. Cyberspace exploitation is often clandestine and includes military intelligence, manoeuvre, information collection, and other enabling actions in preparation for future operations. In other words, it covers intelligence, surveillance and reconnaissance (ISR), espionage, and the planting of APTs. By contrast, cyberspace attack creates noticeable denial effects in cyberspace or manipulates red cyberspace so as to create denial effects in the physical domains. Action verbs define these denial effects, as shown in Table 5.

Table 5. Actions defined in JP 3-12 (US DoD, 2018), p.II-7.

Action	Definition
Deny	To prevent access to, operation of, or availability of a target function by a specified level for a specified time. Deny is sub-divided into Degrade, Disrupt, and Destroy.
Degrade	To deny access to, or operation of, a target to a level represented as a percentage of capacity. Level of degradation is specified. If a specific time is required, it can be specified.
Disrupt	To completely but temporarily deny access to, or operation of, a target for a period of time. A desired start and stop time are normally specified. Disruption can be considered a special case of degradation where the degradation level is 100 percent.
Destroy	To completely and irreparably deny access to, or operation of, a target. Destruction maximizes the time and amount of denial. However, destruction is scoped according to the span of a conflict, since many targets, given enough time and resources, can be reconstituted.
Manipulate	Manipulation, as a form of cyberspace attack, controls or changes information, information systems, and/or networks in grey or red cyberspace to create physical denial effects, using deception, decoying, conditioning, spoofing, falsification, and other similar techniques. It uses an adversary's information resources for friendly purposes, to create denial effects not immediately apparent in cyberspace. The targeted network may appear to operate normally until secondary or tertiary effects, including physical effects, reveal evidence of the logical first-order effect.

3.3 Countering threat networks

Since the targets of OCO will invariably be networks, we also considered JP 3-25 *Countering Threat Networks* (US DoD, 2016). In JP 3-25, threat networks are organisational networks, such as the organisational structure of and communication flow in terrorist and criminal organisations. Threat networks can be engaged using both physical and cyber action. Since action verbs were found in JP 3-25, we included the publication in our analysis.

JP 3-25 (US DoD, 2016) has six chapters. Chapter 5 is most relevant to this paper because it describes the activities required to counter threat networks, including the desired effects (a.k.a. action verbs). The other chapters are not relevant, describing threat networks (Chapters 1 and 2), their environment (Chapter 3), and how to plan operations to counter them (Chapter 4), and how to assess the results (Chapter 6).

There are four strategies for countering threat networks. The counter-resource strategy progressively weakens the threat's ability to conduct operations. The decapitation strategy removes leadership nodes within the network. The fragmentation strategy surgically removes key communication nodes within the network so as to disrupt its ability to function. Finally, the counter-messaging strategy projects messages designed to discourage the recruitment of new members to the threat network and to encourage existing members to leave the network.

The desired effects are selected based on the commander's vision on the future conditions for the threat network and its environment. Action verbs and their definitions are shown in Table 6.

Table 6. Actions defined in JP 3-25 (US DoD, 2016), p.V-4 to V-8.

Action	Definition
Neutralise	To render enemy personnel or materiel incapable of interfering with a friendly operation. Neutralisation of an entire network may not be feasible, but through analysis the commander's staff has the ability to identify key parts to target that will neutralise specific functions.
Degrade	To reduce the effectiveness of efficiency of the threat network by eliminating critical capabilities of the network.
Disrupt	To upset an enemy's formation or tempo, to interrupt the enemy's timetable, or to cause the enemy to commit prematurely or attack in a piecemeal fashion. Should disruption result in the elimination of nodes, the staff must also consider the network's means and time necessary to reconstitute.
Destroy	To physically render an enemy force combat ineffective until it is reconstituted. Alternatively, to damage the enemy so badly that it cannot perform any function or be restored to a usable condition without being entirely rebuilt.
Defeat	To affect the threat network to the extent that it has temporarily or permanently lost the physical means or the will to fight.
Deny	To hinder or deny the enemy the use of territory, personnel, or facilities by destruction, removal, contamination, or obstruction.
Divert	To draw the threat network's attention and forces away from a friendly operation by means of attack, alarm, or feint. Diversions can also cause more circuitous routing, resulting in delays for enemy forces.

4. Analysis

The purpose of this section is to compare the action verbs found in the US DoD doctrine publications with the 6Ds and then to map them to the action verbs found in the US DoD Dictionary.

4.1 Comparing action verbs in doctrine with 6Ds

First, in Table 7 we compare the action verbs obtained from the doctrine publications against the 6Ds. The action verbs are listed top to bottom in alphabetical order.

Table 7. Comparing action verbs with 6Ds.

Cyber kill chain (6Ds)	IO (JP 3-13)	CO (JP 3-12)	CTN (JP 3-25)
Deceive	Deceive	(Deceive – see Manipulate)	
			Defeat
Degrade	Degrade	Degrade	Degrade
Deny	Deny	Deny	Deny
Destroy	Destroy	Destroy	Destroy
Detect	Detect		
Disrupt	Disrupt	Disrupt	Disrupt
			Divert
	Exploit		
	Influence		
		Manipulate (Using deception, decoying, conditioning, spoofing, and falsification)	
			Neutralise
	Protect		
	Respond		
	Restore		

This comparison suggests that the action verbs *degrade*, *deny*, *destroy*, and *disrupt* are common across all three doctrine publications. However, the definitions of the action verbs differ. This is because the intended targets have differing characteristics. In IO, the targets are people's minds, i.e. entities in the *people* layer. In CO, the targets are ICT infrastructure and data, i.e. entities in the *real* / *physical* layer (e.g. hardware), *network* layer, *information* layer (software and data), and possibly *persona* layer (accounts). In CTN, the targets are organisations (in the *social* layer), members and potential members (in the *people* layer), and their communication structures (in the *information* and *network* layers).

4.2 Mapping to verbs in US DoD Dictionary

One way of resolving the differences in definitions is to map the action verbs to those listed in the US DoD Dictionary. Like JP 3-12 and JP 3-04, the latest version (November 2022) of this Dictionary is not currently available on the US Chiefs of Staff Doctrine website. A search of the Internet showed that the most up-to-date version (November 2021) could be found on the Federation of American Scientists' website (US DoD, 2021). This version lists the action verbs as follows (ibid., p.2):

> *"The following is a lengthy but not whole inclusive list of strategic effect terms: advance, assure, coerce, compete, compel, contain, deceive, defeat, degrade, delay, delegitimize, deny, destroy, deter, discredit, disable, discourage, disrupt, divert, engage, enhance, integrate, isolate, kill, maintain, manage, neutralize, prevent, protect, stabilize, suppress, synchronize."*

Unfortunately, only one of these 32 verbs (*neutralize*) is defined in the DoD Dictionary. Moreover, there were 11 more verbs in the pilot study and the doctrine publications. Since the DoD Dictionary states (US DoD, 2021, p.2): "*For strategic effect terms, the standard dictionary definition often applies"*, we decided to follow this suggestion for all 43 verbs. For convenience, we chose the online Collins English dictionary (https://www.collinsdictionary.com/dictionary/english).

The full list was too unwieldy for our research. We observed that several verbs described action on a situation or condition, rather than action on an entity (such as those listed in Table 1). In addition, there were several verbs that could be regarded as synonyms. After pruning, the resulting action verbs, together with their source, any synonyms, and the definition selected from the Collins English dictionary, are listed in Table 8. The original 6Ds are underlined.

Table 8. Action verbs resulting from mapping to DoD Dictionary definitions (6Ds underlined).

Verb	Source(s)	Synonym(s)	Definition (Collins)
coerce	Dictionary, JP 3-13, JP 3-12	compel, deter, discourage, influence, condition	2. If you coerce someone into doing something, you make them do it, although they do not want to.
deceive	Dictionary, 6Ds, JP 3-12	decoy, spoof, falsify	If you deceive someone, you make them believe something that is not true, usually in order to get some advantage for yourself.
degrade	Dictionary, 6Ds		2. To degrade something means to cause it to get worse.
delay	Dictionary, pilot	latency	2. To delay someone or something means to make them late or to slow them down.
deny	Dictionary, 6Ds, JP 3-13	contain, isolate, prevent, protect, suppress	3. If you deny someone something that they need or want, you refuse to let them have it.
destroy	Dictionary, 6Ds	kill	1. To destroy something means to cause so much damage to it that it is completely ruined or does not exist anymore.
detect	6Ds		1. To detect something means to find it or discover that it is present somewhere by using equipment or making an investigation. 2. If you detect something, you notice it or sense it, even though it is not very obvious.
disrupt	Dictionary, 6Ds	disable, neutralize	If someone or something disrupts an event, system, or process, they cause difficulties that prevent it from continuing or operating in a normal way.
divert	Dictionary, JP 3-25		1. To divert vehicles or travellers means to make them follow a different route or go to a different destination than they originally intended.
exploit	JP 3-13	manipulate	3. If you exploit something, you use it well, and achieve something or gain an advantage from it.
identify	(See text)	recognize, classify	1. If you can identify someone or something, you are able to recognize them or distinguish them from others.
integrate	Dictionary		3. If you integrate one thing with another, or one thing integrates with another, the two things become closely linked or form part of a whole idea or system.
move	Dictionary	advance	2. When you move, you change your position or go to a different place.
observe	(See text)	Monitor, survey	1. If you observe a person or thing, you watch them carefully, especially in order to learn something about them.
penetrate	(See text)		1. If something or someone penetrates a physical object or an area, they succeed in getting into it or passing through it.
respond	JP 3-13	react	1. When you respond to something that is done or said, you react to it by doing or saying something yourself.
restore	Dictionary, JP 3-13	maintain	2. To restore someone or something to a previous condition means to cause them to be in that condition once again.
seize	pilot	capture	3. If a government or other authority seize someone's property, they take it from them, often by force.
synchronize	Dictionary		If you synchronize two activities, processes, or movements, or if you synchronize one activity, process, or movement with another, you cause them to happen at the same time and speed as each other.

The author's choice of definitions, of which verbs to retain, and of which to regard as synonyms is unavoidably subjective. Some verbs require explanation, as follows:

- *Observe* and *identify*. Neither the doctrine publications nor the DoD Dictionary include terms that describe the ISR process. The 6Ds only hint at ISR in the guise of *detect*. Detection of potential targets occurs as a result of observing an area, and after detection the target is identified as an instance of a particular entity and classified as friendly, hostile, or neutral.
- *Move*. The first term in the DoD Dictionary list is *advance*, but there are no other terms denoting movement in physical or cyber space. We considered the term *manoeuvre*, but rejected this because it is a more complex concept.
- *Penetrate*. In OCO, penetration of the target system is a key step in the attack process. In effect, penetrate is overcoming the defender's *deny* (a.k.a. *protect*) action.
- *Seize*. The term *seize*, together with its synonym *capture*, was identified in the pilot study. Although it was only used for the physical domain, the verb is equally applicable to taking (over) control of cyber entities.
- *Respond*. The majority of action verbs found are proactive in nature. Even so, some actions will be a reaction to the defender's actions. We adopted the verb *respond* from JP 3-13.

5. Findings

While comparing the action verbs and mapping them to the verbs in the DoD Dictionary, we observed a number of conceptual shortcomings in the 6Ds. The lack of verbs fully describing the ISR process has already been mentioned. Other shortcomings are as follows:

- *Space, location, distance, and movement*. The 6Ds lack concepts of location, of distance between locations, and of movement. Both physical and cyber action take place in a space. Physical space is continuous and has three dimensions. Cyberspace is discrete and multi-dimensional. For example, an important aspect of OCO is gaining access to the target system (Anon & Anon, 2022), The access path may cross both physical and cyber space (cf. over an air-gap, as in Stuxnet).
- *Time, duration, and synchronization*. The need to add the action verbs *delay* and *synchronize* show that the 6Ds lack concepts of time, duration, and synchronization.
- *Control over entities*. The need to add the action verb *seize* showed that the 6Ds lacked the concept of the ownership or control over entities.
- *Aggregation and disaggregation of entities*. The action verb *integrate* from the DoD Dictionary describes aggregating several individual entities to become a single compound entity. The reverse process splits a compound entity into its constituent entities. The 6Ds do not cover these concepts.

In summary, the 19 action verbs listed in Table 8 can be grouped into the following categories:

- Spatial verbs: *divert*, *move*.
- Temporal verbs: *delay*, *synchronize*.
- ISR verbs: *detect*, *identify*, *observe*.
- Entity aggregation verbs: *integrate*.
- Entity control verbs: *seize*.
- Effects on targets: *coerce*, *deceive*, *degrade*, *deny*, *destroy*, *disrupt*, *exploit*, *penetrate*, *respond*, *restore*.

Finally, we can answer the questions we set ourselves, as follows:

- Are the 6Ds applicable to OCO? Yes, as demonstrated by JP 3-12.
- Are some verbs preferred for cyber action and others for kinetic action? No, we found no evidence for preferences. All the verbs in Table 8 can be applied equally well to physical and cyber entities, perhaps using synonyms (e.g. *pivot* for *move)*.
- Which is the greatest of the 6Ds? No evidence could be found for any one being the greatest.

6. Conclusions and Further Research

The cyber kill chain approach (Croom, 2010) introduced a proactive strategy for defending computer-based systems from intrusion. The defender can interrupt the intruder's kill chain in any phase by taking appropriate action. Drawing on US Department of Defense doctrine for information operations, Hutchins, Cloppert & Amin (2011) identified the defender's possible actions as *detect*, *deny*, *degrade*, *disrupt*, *destroy*, and *deceive* (a.k.a. "the 6Ds").

This paper identifies which action verbs are best suited to OCO in the context of all-domain operations (ADO). The focus is on action by professional organisations at the operational and tactical levels. A pilot study (Grant & Kantola, 2021) suggested that *deceive*, *delay*, and *deny* were suited to cyber action and *seize*, *capture*, and *destroy* to physical action. It identified more verbs than the 6Ds.

The 6Ds were traced back to the 2006 version of the US DoD Joint Publication (JP) 3-13 *Information Operations*. This publication included yet more verbs. When cyberspace operations were split off from information operations around 2010, separate doctrine was published as JP 3-12 *Cyberspace Operations*. This document included five of the 6Ds, solely in OCO. Four of the 6Ds were also found in JP 3-25 *Countering Threat Networks*. These observations confirmed that the 6Ds were applicable to OCO as well as defence.

All the action verbs found were mapped to the 32 strategic effect terms found in the US DoD Dictionary of Military and Associated Terms. After pruning out synonyms and action on situations, the remaining verbs exhibited shortcomings in that they lacked any representation of space, time, ISR, and entity control and aggregation. Rectifying these shortcomings yielded a set of 19 action verbs for which definitions were obtained from the Collins English dictionary. All 19 could be applied to both cyber and physical action.

The main contribution of this paper is that it harmonizes action verbs from a variety of OCO-related doctrine publications with definitions from a standard English dictionary. In doing so, it rectifies shortcomings, groups them into categories, and shows that all verbs are applicable to ADO. The main limitation is that the choice of definitions and of which verbs to retain is subjective.

Further work is needed to relate the action verbs to attack phases (as in Table 3) and to entities in the six layers of physical and cyber space (see Table 1). Moreover, it may be possible to identify common sequences of action verbs (e.g. *move*, *observe*, *detect*, *identify*, *exploit*), together with the entities involved. More fundamentally, theory and doctrine could be linked by revisiting older work on deriving action verbs from Boisot's model of data, information, and knowledge (Hutchinson & Warren, 2001) and from the Shannon-Weaver model of communications and hypergame theory (Kopp, 2003).

References

Anon & Anon. (2022) All About Access: Inzichten en implicaties van MIVD-cyberoperaties voor digitale slagkracht. *Militaire Spectator*, 191, 9, 464-475. [In Dutch: All About Access: Insights and implications of MIVD cyber operations for digital fighting power.]

Borne, D. (2019). Targeting in Multi-Domain Operations. *Military Review*, May-June, 60-67.

Croom, C. (2010). The Cyber Kill Chain: A foundation for a new cyber security strategy. *High Frontier: The journal for space and cyberspace professionals*, 6, 4, 52-56, August 2010, US Air Force Space Command.

Denning, P.J. & Denning, D.E. (2010). Discussing cyber attack. *Communications of the ACM*, 53, 9, 29-31.

Grant, T.J. and Kantola, H. (2021). Targeting in All-Domain Operations: choosing between cyber and kinetic action. Proceedings, 20th European Conference on Cyber Warfare and Security, 24-25 June 2021, 139-148.

Hutchins, E., Cloppert, M. & Amin, R. (2011). Intelligence-driven Computer Network Defense informed by Analysis of Adversary Campaigns and Intrusion Kill Chains. Proceedings, 6th International Conference of Information Warfare, 17-18 March 2011, 113-125.

Hutchinson, W. & Warren, M. (2001). *Information Warfare: Corporate attack and defense in a digital world*. Routledge, Abingdon, UK.

Jackson, A.P. (2013). *The Roots of Military Doctrine: Change and continuity in understanding the practice of warfare*. Combat Studies Institute Press, Fort Leavenworth, Kansas, US.

Kopp, C., 2003. Shannon, hypergames and information warfare. *Journal of Information Warfare*, *2*(2), 108-118.

Lin, H. (2009). Lifting the veil on cyber offense. *IEEE Security & Privacy*, 7, 4, 15-21.

Pomerleau, M. (2022). DOD publishes revised doctrine on information. 7 October 2022, DefenseScoop, https://defensescoop.com/2022/10/07/dod-publishes-revised-doctrine-on-information/ (accessed 26 October 2022).

Theohary, C.A. (2022). Defense Primer: Information Operations, In Focus, Library of Congress, Congressional Research Services report IF10771, version 9, 9 December 2022.

UK MoD. (2016). *Cyber Primer*. 2nd edition, Development, Concepts & Doctrine Centre, MoD Shrivenham, UK, July 2016.

US DoD. (2006). *Information Operations*. US DoD Joint Publication 3-13, 13 February 2006.

US DoD. (2016). *Countering Threat Networks*. US DoD Joint Publication 3-25, 21 December 2016.

US DoD. (2018). *Cyberspace Operations*. US DoD Joint Publication 3-12, 8 June 2018.

US DoD. (2021). *US DoD Dictionary of Military and Associated Terms*. November 2021. https://irp.fas.org/doddir/dod/dictionary.pdf (accessed 17 December 2022).

Known Unknowns: The Inevitability of Cyber Attacks

Virginia A. Greiman
Boston University, Boston, MA, USA
ggreiman@bu.edu

Abstract: As described by Former U.S. Secretary of Defense, Donald Rumsfeld in his 2011 book, Known and Unknown, "there are many things of which we are completely unaware—in fact, there are things of which we are so unaware, we don't even know we are unaware of them." Throughout history the world has faced numerous catastrophic events that were not foreseen but in hindsight were discoverable including the devastating effects of Pearl Harbor, and the September 11 terrorist attacks. More recently, the potential for catastrophic loss has been magnified in the 2020 Solar Winds and 2021 Colonial Pipeline cyber-attacks. We may not know when or how these events will occur or how much damage or destruction will occur, but we do know that these events are possible. The literature differentiates between events that occur totally by surprise, and outcomes or events that actors have identified as possibly existing but do not know whether they will take place or not. The aim of this paper is to provide insight, based on an empirical review of selected attacks both within and outside the cyber space literature to uncover the underlying risk, uncertainty, and complexity that may have been known but not seriously considered by those who had the knowledge and capability to investigate the warning signs. Based on the case study analysis, this paper will present the reasons for inaction and how we can learn from these experiences. The following two theories – institutionalization and rationalization have been found to provide some reasons for the occurrence of behaviors which increase the possibility of unobserved risks. In this paper we explore these theories through case study analysis and propose a framework consisting of four concepts for increasing awareness of these situations.

Keywords: Known Unknown, Cyber Attacks, Cyber Security Risk, Unobserved Cyber Risk, Risk Perception

1. Introduction

Notably, according to the Center for Strategic and International Studies, since 2006 there have been more than 900 major global cyber incidents that had potential for serious catastrophic loss (CSIS, 2023). In the United States, many Members of Congress have raised concerns over the frequency, types, and impacts of cyber incidents during hearings, speeches, and in legislation (Jaikaran, 2021). Cyber incidents affect nearly every national entity, from federal and state government agencies to non-profits, private companies and individuals.

The Director of National Intelligence is required annually to deliver to Congress an assessment from the intelligence community on worldwide threats. Recent assessments have highlighted cyberspace as an area of strategic concern, with Russia, China, Iran, and North Korea as the leading threat actors. Attacks from these countries include spying on government agencies by accessing agency computers, stealing sensitive information from public and private sector entities in the United States, and destroying or potentially destroying computer equipment (Jaikaran 2021).

In March 2022, The U.S. Justice Department unsealed charges accusing four Russian officials of carrying out a series of cyberattacks targeting critical infrastructure in the United States, including a nuclear power plant in Kansas, and evidently compromising a petrochemical facility in Saudi Arabia. The announcement covered hackings from 2012 to 2018, but served as yet another warning from the Biden administration of Russia's ability to conduct such operations (DOJ 2022).

The frequency, type, and impact of cyber incidents against and in the United States continues to grow (Statista 2021). In an effort to address this challenge, policymakers are considering a variety of solutions, such as denying opportunities for successful attacks by improving defenses and deterring adversaries from engaging in disruptive activities in cyberspace. In this research we introduce another solution: finding better ways to identify and respond to the unknowns.

To guide this research four cases studies were selected involving significant attacks both within and outside the cyber-attack literature and were investigated through a review of congressional hearings and witness statements to determine the specific context of an event and the response from the perspectives of the participants (Mills, Durepos and Wiebe 2010). The small number of cases selected reflects a desire to take a deeper look at each case rather than a quantitative analysis of a large number of cases.

1.1 Exploring the Unknown

Based on some of the most significant events of our time including the horrific attack on Pearl Harbor, the September 11 terrorist attack on the United States and the recent unprovoked attacks in Ukraine it is critical to

reduce uncertainty in our complex and unknown world if we are going to prevent repeating the catastrophic losses of the past. According to the Congressional Report on Pearl Harbor, early warning signs were neglected by the United States hours before the attack (U.S. Senate 1946). Because of the surprise factor, the potential that a weaker nation might use such an attack to target a stronger one and other potential similarities, these attacks have been commonly referred to as cyber Pearl Harbor attacks (Straub 2021).

In its report on the terrorist attacks of Sept. 11, 2001, the 9/11 Commission noted that the intelligence community, assailed by "an overwhelming number of priorities, flat budgets, an outmoded structure, and bureaucratic rivalries," had failed to pin down the big-picture threat posed by "transnational terrorism" throughout the 1990s and up to 9/11. In response to the 9/11 Commission's recommendations, Congress created a national intelligence director and the National Counterterrorism Center to pool intelligence.

Intelligence failures are commonly understood as the failures to anticipate important information and events, such as terrorist attacks. However, intelligence can fail for many reasons, often despite the best work of intelligence professionals (Dahl 2013). Explanations for intelligence failure generally include one of more of the following causal factors: organizational obstacles, psychological and analytical challenges, problems with warning information, and failures of political leadership (Copeland 2017).

Understanding the decision making behaviors that produce successful outcomes in the recognition and identification of known/unknown risk factors is essential to preventing and reducing the risk of catastrophic cyber-attacks. The serious problem of cyber-attack and cybersecurity means finding the "unknown unknowns" (MIT 2021). Former U.S. Secretary of Defense, Donald Rumsfeld, distinguishes between categories of risk (Rumsfeld 2011). These categories described below are recognized in the defense industry, risk management, and the cyber security literature among other disciplines though not always described in the same way.

1.2 Known Knowns

The first category is identified as "known knowns" describing the things we know we know. Examples could be an organization's location, the individuals that are impacted, and other realities. Known knowns are assumptions that have been validated and are verifiable. However, it is important to recognize that known knowns need to be continually evaluated for change. For example, on the 12th of February 2002, Donald Rumsfeld, used a little known framework to help him in making the case for the invasion of Iraq. The known and unknown framework may tell us that a country has no weapons of mass destruction on a particular day, but a week later that may change.

1.3 Known Unknowns

The second category is defined as "known unknowns" describing the things we know can happen but we are uncertain of the results. Much scientific research is based on investigating known unknowns. In other words, scientists develop a hypothesis to be tested, and then in an ideal situation experiments are best designed to test the null hypothesis (Logan 2009, p. 712). Known unknowns are usually listed in an organization's risk register and mitigated to the extent possible. For instance, we may be aware of certain possible events such as the existence of hurricanes and weather change in certain locations and at certain times of the year, but we don't know what the impact might be and when and where the damage may occur. As humans, we know we are mortal but we don't know how long we will live. Or, we know we may have economic downturns in the future, but we don't know when and for how long they may occur.

1.4 Unknown Unknowns

The third category is defined as the "unknown unknowns". This category describes uncertainties that we could not have known in advance and let alone foresee their consequences, e.g., natural cataclysms. Unknown unknowns can arise from lack of awareness, insufficient information or knowledge, and from organizational, psychological, political, or cultural blindness. These risks often cannot be identified precisely due to multiplicity, but whose total negative impact appears certain. These include an act of terrorism, criminal acts, a depression or revenue risk. We may not know when or how these events will occur or how much damage or destruction will result from the event (unknown), but we do know that all of these events are possible.

1.5 Unknown Knowns

As described by Donald Rumsfeld, these are things we know, but don't know we know. Often, they are uncovered after the fact. For example, we may know that every major cybersecurity incident is a wake-up call for the nation, (known) however, no decisions were made on this knowledge as the immediate memories of the incident fade. We know that in the cyber world supply chain attacks can occur (known) but we may not have investigated our own supply chain to know how serious the problem may be (unknown). Unknown knowns have been found to be commonplace in organizations because we can know things, but not realize how important they are or how they fit together. The Solar Winds attack is a great example of an unknown known.

1.6 Theories of Unobserved Risk

The literature reflects the following theories – institutionalization and rationalization to provide some reasons for the occurrence of behaviors which increase the possibility of unobserved risks. We explore these theories to provide a deeper understanding of how these may impact the ability to uncover the unknown.

1.7 Institutionalization: The 9/11 Terrorist Attacks

The very structure of business and governmental organizations, particularly those that are large and complex, makes it difficult to anticipate predictable surprises. Because companies are usually divided into organizational silos, the information leaders need to see and assess an approaching threat is often fragmented (Watkins and Bazerman 2003). Various people have various pieces of the puzzle, but no one has them all. And those at the top inevitably receive incomplete and distorted data. On September 11, 2001 various government agencies had pieces of information on terrorists' methods and plans that, had they been combined, would have pointed to the type of attack that was carried out against the World Trade Center and the Pentagon. Tragically, the information remained fragmented.

According to the 9/11 Report, "most of the intelligence community recognized in the summer of 2001 that the number and severity of threat reports were unprecedented. "Many officials told us that they knew something terrible was planned, and they were desperate to stop it. Despite their large number, the threats received contained few specifics regarding time, place, method, or target. Most suggested that attacks were planned against targets overseas; others indicated threats against unspecified 'U.S. interests.'" We cannot say for certain whether these reports as dramatic as they were, related to the 9/11 attacks (Kean and Hilton 2004, pp. 262-263).

1.8 Rationalization: The Ariane 5 Rocket Disaster

Rationalization enables system operators to convince themselves that not responding to a risk or threat is not only legitimate, but acceptable and perhaps even necessary. The justification for deviating from a standard, process or procedure includes the belief that the procedure is unnecessary or the work environment is dynamic, unstable and unpredictable.

On June 4, 1996, the massive Ariane 5 rocket, launched on its maiden voyage from the European Space Agency's center in French Guiana, exploded 39 seconds into flight, destroying four satellites on board (ESA, 1996). At an altitude of about 3700 m, the launcher veered off its flight path, broke up and exploded. According to the Inquiry Board Report: The failure of the Ariane 501 was caused by the complete loss of guidance and altitude information 37 seconds after start of the main engine ignition sequence (30 seconds after lift-off).

In a postmortem analysis, the European Space Agency (ESA) noted that the problem was due to a malfunction in the rocket's guidance system software, and it acknowledged that the system was not fully analyzed or understood, perhaps because the software had worked successfully with Ariane 4 (Ramesh and Browning 2014) and the participants rationalized it would work again. More generally, the disastrous outcome was due to an uncertainty of which the Agency was unaware prior to the launch— (an unknown unknown). Yet was this uncertainty truly unknowable, or was it potentially knowable but just escaped recognition from the project team? (Greiman 2023)

1.9 Case Studies

For purposes of this research four case studies were selected shown in Table 1 to understand the circumstances that led to two of the worse disasters in the history of mankind – The devastating attack on Pearl Harbor on

December 7, 1941 - and the terrorist attack on the World Trade Center on 9/11. The Solar Winds Attack in 2020 and the Colonial Pipeline attack in 2021 were selected to demonstrate the catastrophic potential of a cyber-attack when warnings are not heeded. Each case is analyzed based on the following five factors (1) the event or threat; (2) the warnings provided; (3) the reasons attributed to the failure to respond; (4) the impact of the event; and (5) the attribution.

Table 1: Case Studies on Responses to Known and Unknown Events

Event/Threat	Warnings	Failure to respond	Impact	Attribution
Pearl Harbor December 7, 1941	Japanese submarine sunk in the harbor; Japanese patrol planes spotted on radar but disregarded. The Japanese Armada of attack planes were mistakenly identified as a fleet of U.S. bombers (U.S. Senate 1946).	Gross incompetence or outright dereliction of duty on the part of the officers and officials in charge (U.S. Senate 1946); Background of noise, extraneous or misleading information that can obfuscate emerging events (Wohlstetter 1962)	The attack on Pearl Harbor took the lives of more than 2,400 Americans and sunk or damaged 21 ships in the U.S. Pacific fleet (U.S. Senate 1946).	"The ultimate responsibility for the attack and its results rests upon Japan," the final congressional committee report concluded, and "the diplomatic policies and actions of the United States provided no justifiable provocation whatever for the attack by Japan on this Nation" (U.S. Senate 1946)
September 11 Terrorist Attacks	The system was blinking red (Kean and Hamilton 2004, n. 30, p. 517). Various government agencies had pieces of information on terrorists' methods and plans that, had they been combined, would have pointed to the type of attack that was carried out against the World Trade Center and the Pentagon. Tragically, the information remained fragmented (Kean and Hamilton 2004).	"The September 11 attacks fell into the void between the foreign and domestic threats. The foreign intelligence agencies were watching for foreign threats to U.S. interests there. The domestic agencies were waiting for evidence of a domestic threat from sleeper cells within the U.S. No one was looking for a foreign threat to domestic targets. The threat that was coming was not from sleeper cells. It was foreign-but had come from foreigners who had infiltrated into the United States" (Kean and Hamilton 2004, p. 63).	September 11 took the lives of 2,977 innocent people. The United States' response to the 9/11 attack, namely the invasion of Afghanistan, had a deeply destabilizing impact on the Middle East and North Africa region. In addition to the death of tens of thousands of Afghanis and others from the region, as well as the emergence of new terrorist groups.	Invasion of Iraq and Afghanistan; however, the threat of Taliban retaliations still exist today.
2021 Colonial Pipeline Ransomware	Attacks on critical infrastructure was entirely predictable when Stuxnet was discovered in 2010. Refusal to participate in physical and cyber pipeline security TAS assessments showed complacency and over confidence (Congress 2021).	Response plan did not include ransomware – even though ransomware attackers had been targeting critical infrastructure since 2015 (Congress 2021)	The attack on May 7 resulted in a week-long shutdown of 5,500 miles of petroleum pipeline on the East Coast that clearly represents a significant cyber-attack on critical transportation infrastructure	The lack of a global agreement on legal attributions of criminal responsibility portends a continuation of cyber-attacks until an agreement is reached.
2020 Solar Wind Hack Espionage-based assault	Prior to and throughout the long period of attack there were allegations of missed opportunities, hints of problems that were ignored, lax security and the failure of U.S. intelligence officials to connect the dots of the supply chain attack and the insertion of the	As one expert expressed, there was not enough detail to report the problem to SolarWinds or the U.S. government. "We thought we didn't have enough evidence to reach out," (Temple-Raston 2021). Critics said they should have seen the hackers	Broad attack on confidential information of U.S. Government including Homeland Security and the Pentagon and the tech tools used by companies to protect them. SolarWinds estimates that 18,000 of its over 300,000 customers are vulnerable to this malware. The U.S.	Alleged Russian Hacker Group. In April 2021 President Biden ordered a new round of economic sanctions on Russia — a response in part to a Kremlin-linked computer breach that penetrated numerous U.S. government networks.

Event/Threat	Warnings	Failure to respond	Impact	Attribution
	malicious code (Temple-Raston 2021).	from the Russian intelligence service, the SVR, preparing this attack (U.S. Senate 2021).	government estimates that as many as 250 government agencies and companies may have been affected (U.S. Senate 2021).	

1.10 Reducing the Uncertainty Surrounding Cyber Events

Due to the page limitation of this paper, we cannot fully explore the reasons behind the failure of individuals to respond to warnings of potential cyber threats or the theories behind these failures, and will reserve this analysis for future research with a larger number of case examples. Instead, in the balance of this paper we will focus on solutions to reducing the potential for unknowns by creating greater awareness of the threat environment. Thus, reducing the likelihood that institutionalization of rationalization will interfere with the opportunity to recognize the potential for a cyber-attack. The theory being that by reducing uncertainty, the potential victims will be better prepared to respond to the unknown, especially when cyber-attacks cannot be fully prevented or predicted. The framework and concepts below are intended to explore the various ways in which organizations can address uncertainty and limit the likelihood that warning signs will be ignored. through discovery-driven planning or by systemically probing for knowledge gaps

1.11 Framework and Concepts for Responding to the Unknown

The following four concepts were derived from the literature on dealing with the unknown in complex, intractable and sometimes impossible situations: (1) Accountability at the international level; (2) Reducing uncertainty; (3) Mindful organizing and (4) Sensemaking.

These concepts have emerged from research and actual experience from both failed and successful outcomes from unexpected attacks both within and outside the cyber-attack literature particularly when faced with significant or emergent events with an element of surprise. Though not all of these concepts will be beneficial in all circumstances, the goal is to encourage reflection and stimulate ideas for approaching the unknown and uncertainty that lies in today's complex global domain of cyberspace.

These concepts are summarized in the framework shown in Figure 1.

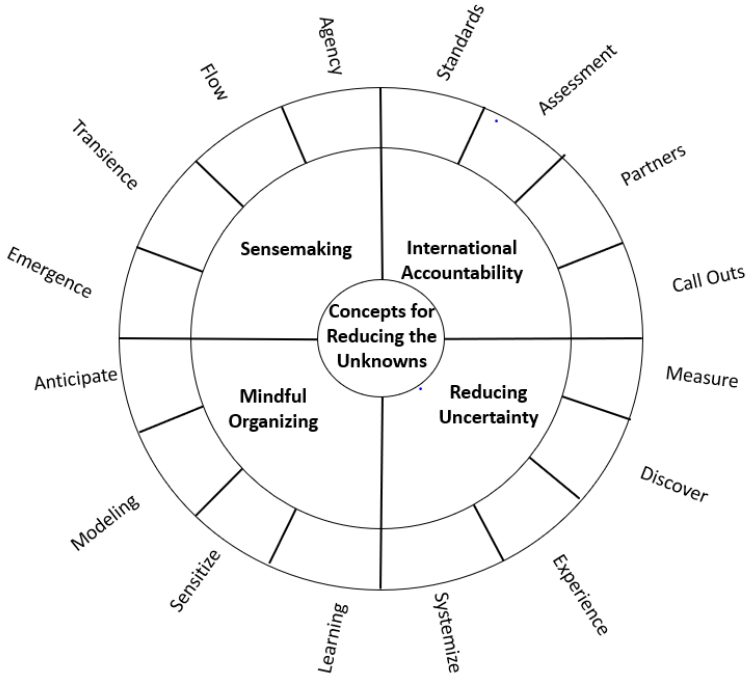


Figure 1: Concepts for Reducing the Unknown

2. International Accountability

National government and intergovernmental organizations as well as legal scholars have noted the need for accountability at the international level when one country is wrongfully attacked by another (Kuntz 2013; CEEAA 2014). Scholars recognize that States must develop accountable attribution mechanisms for international law to have practical value in the cyber sphere (Banks 2019).

An effective way of enhancing international accountability may be the establishment of standards or norms as proposed by NATO, the OECD, and other transnational organizations. “This requires a calling out of wrongful acts conducted by other states, through a harmonized attribution process, something that victimized states have been reluctant to do” (Hathaway 2017, p. 5). Moreover, because states are unable to affect the risk calculus of other states through conventional deterrence mechanisms, the next best thing is to foster an environment in which states consistently interact to provide information to one another (Brantly 2021).

Several non-state-driven norms-making initiatives have sought to fill the void, including Microsoft's (2016) cyber norms proposals, the Tallinn Manual project (Tallinn 2017), and the Global Commission on the Stability of Cyber Space (GCSC) (Eggenschwiler 2020). Scholars have contended that this emerging body of non-binding norms presents states with a critical window of opportunity to reclaim a central law-making position, similar to historical precedents including the development of legal regimes for Antarctica and nuclear safety (Mačák 2017).

Researchers have proposed a due diligence framework or minimal standards for attribution of cyber-attacks (Chircop 2018). Though due diligence is not clearly defined in the law it was reflected by the International Court of Justice (ICJ) in its Corfu Channel Judgment: “It is every State’s obligation not to allow knowingly its territory to be used for acts contrary to the rights of other states” (Corfu 1949) and has been set forth as an obligation of States in the pronouncements of the International Law Association (ILA 2014).

According to the Defense Science Board (DSB 2017) as offensive cyber capabilities continue to grow, and are likely to outpace cyber defense, there are likely to be growing risks of misperception that could lead to rapid cyber escalation – and the potential for rapid escalation to armed conflict requiring the necessity of working closely with the private sector.

3. Reducing Uncertainty

Lenfle and Loch (2017) in their studies of megaproject failure found that the first of the three major causes of failure is: “Underestimation of or refusal to acknowledge uncertainty.” Thus, an increase in complexity often means that it is more difficult to comprehend the effect of influencing one element; hence there is increased uncertainty. Put differently, uncertainty refers to the components, relationships, and interactions we do not fully comprehend or of which we may not even be aware. Complexity and uncertainty are thus strongly related (Salet 2021).

Finding ways to reduce the unknown is a central focus of this research. Thus, reducing uncertainty will contribute to the goal of reducing surprise and preventing potential cyber disasters. Experience has found that uncertainty that is unforeseen can be diagnosed through discovery-driven planning or by systemically probing for knowledge gaps (McGrath and McMillan 2000). Uncertainty refers to the possibility of emergent indeterminate events. Traditionally, organizations particularly in cyber security have focused on what economists call risk and have not tended to deal explicitly with uncertain, emergent, and unforeseen events (Miller and Hobbs 2005).

4. Mindful Organizing

The wisdom of learning from failure is incontrovertible. Yet organizations that do it well are extraordinarily rare. High-reliability-organizations (HROs) are an example of organizations that have developed practices to help prevent catastrophic failures in complex systems like nuclear power plants aircraft carriers through early detection. According to Weick & Sutcliffe (2015) HROs demonstrate particular characteristics in the way they operate: anticipating problems (being aware of what is happening in the work system; being alert to ways in which an incident could occur; looking beyond simplistic explanations for incidents; and containing problems using relevant expertise regardless of where it is situated within the organizational hierarchy). They identified five principles they define as “mindful organizing” that include preoccupation with failure, a reluctance to simplify, a sensitivity to operations, a commitment to resilience and a deference to expertise as a shared set of values that foster resilience through constant communication and recalibration in the face of unknowable risks (Weick & Sutcliffe 2015). High reliability theory should be explored in the cyber field because of the high potential

for catastrophic outcomes from a cyber incidence. For instance, a cyber-attack on a nuclear plant could have long lasting and deadly consequences.

5. Sensemaking

Sensemaking, a term introduced by Karl Weick, refers to how we structure the unknown so as to be able to act on it and has been defined as a key leadership capability for the complex and dynamic world we live in today (Ancona 2011). Sensemaking involves producing a plausible understanding—a map—of a shifting world; testing this map with others through data collection, action, and conversation; and then refining, or abandoning, the map depending on its credibility.

As described by Weick, Sutcliff and Obtsfeld (2005) to focus on sensemaking is to portray organizing as the experience of being thrown into an ongoing, unknowable, unpredictable streaming of experience in search of answers to the question, "what's the story?" (Weick et al 2005). Plausible stories animate and gain their validity from subsequent activity. The language of sensemaking captures the realities of agency, flow, equivocality, transience, re-accomplishment, unfolding, and emergence, realities that are often obscured by the language of variables, quantities, and structures (Weick et al 2005). Students of sensemaking understand that the order in organizational life comes just as much from the subtle, the small, the relational, the oral, the particular, and the momentary as it does from the conspicuous, the large, the substantive, the written, the general, and the sustained (Weick et al 2005). The concept of sensemaking fills important gaps in organizational theory.

5.1 Future Approaches to Counter Cyber Attacks

The approach taken by most organizations to counter cyber-attacks is defensive and reactionary. Threats are only removed and analyzed once they are detected; at which point, the harm is done—the network has already been breached and valuable information compromised.

A lot of solutions have been offered to prevent or reduce cyber-attack impacts, including mandatory reporting requirements, wider use of multi-factor authentication, requiring a software bill of goods, and significantly improving threat information sharing between the government and the private sector. However, these solutions will not be sufficient in light of the lack of understanding of the large number of unknown unknowns prevalent in the domain of cyber space.

Predictive analytics will have a prominent role in the future of cybersecurity given its use of historical data and statistical algorithms to predict future threats. The most promising aspect of predictive analytics is that with the meteoric rise in machine learning through Artificial Intelligence (AI), it will be possible to teach machines how to do it; through which human error can be eliminated. This will introduce a very high level of efficiency to detecting threats very early on while giving cybersecurity a chance to be proactive instead of reactive.

6. Conclusion:

Cyber security risk in the U.S. and around the world is a major national security threat impacting all nations. Analyzing the underlying causes for failure to respond to these threats can make these surprise attacks less likely. Cyber security failures will no longer be inevitable if we find ways to identify and take appropriate action on the unknowns so that policymakers may gain a greater understanding of the cyber risks that the nation and the world faces. Knowledge of known adversaries, the types of activities they conduct online, and how they are identified by national governments will inform the debate Until we have a greater understanding of the reasons behind the failure to respond to the warning signs of potential cyber-attacks the inevitability of countering cyber-attacks before they occur will remain.

Acknowledgements

I would like to thank my graduate research assistant, Renee Macasaet for the extensive and valuable research she contributed to this paper.

References

Ancona, D. (2011) Sensemaking: Framing and Acting in the Unknown. In: S. Snook, N. Nohria, & R. Khurana, (eds.) (2011). *The Handbook for Teaching Leadership: Knowing, Doing, and Being*. Thousand Oaks, CA: Sage Publications.

Banks, W. (2019) The Bumpy Road to a Meaningful International Law of Cyber Attribution. *American Journal of International Law* Unbound, vol. 113, pp 191-196. doi:10.1017/aju.2019.32.
Brantly, A.F. (2021) Risk and uncertainty can be analyzed in cyberspace. *Journal of Cybersecurity*, 2021, 1–12.
Center for Strategic and International Studies (2022) *Significant Cyber Incidents Since 2006*, CSIS, Washington, D.C.
Chircop, L. (2018) A Due Diligence Standard of Attribution in Cyberspace. *International and Comparative Law Quarterly*, Vol. 67, no. 3, pp 643-668. doi:10.1017/S0020589318000015.
Copeland, T.E. (2017) *Intelligence Failure Theory*. International Studies Association and Oxford University Press, Oxford, England.
Corfu Channel (United Kingdom v Albania) (Judgment) [1949] ICJ Rep 4, 22 (Corfu Channel).
Dahl, E. (2013) *Intelligence and Surprise Attack: Failure and Success from Pearl Harbor to 9/11 and Beyond.* Georgetown University Press, Washington, D.C.
Defense Science Board (DSB) (2017) *Task Force on Cyber Deterrence*. Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics, Washington, D.C.
Eggenschwiler, J. (2020) Expert commissions and norms of responsible behaviour in cyberspace: a review of the activities of the GCSC. *Digital Policy, Regulation and Governance*, Vol. 22 No. 2, pp. 93-107. https://doi.org/10.1108/DPRG-03-2019-0019.
Greiman, V.A. (2023) *Global Megaprojects: Lessons, Case Studies and Expert Advice on International Megaproject Management*. John Wiley & Sons, Hoboken, NJ
Hathaway, M. (2017) *Getting Beyond Norms*, CIGI Papers No. 127, Centre for International Governance Innovation, Ontario, Canada.
International Law Association (ILA) (2014) ILA Study Group on Due Diligence in International Law (First Report, ILA, 7 March).
Jaikaran, C. (2021) Cybersecurity: Selected Cyberattacks, 2012-2021. Congressional Research Services (CRS), Washington, D.C.
Kean T.H., and Hamilton L.H. (2004) *The 9/11 Commission Report: Final Report of the National Commission on Terrorist Attacks upon the United States*. Norton, New York, NY.
Kuntz, R. L. (2013) How Not to Catch A Thief: Why the Economic Espionage Act Fails to Protect American Trade Secrets, 28 *Berkeley Tech. L.J.* 901.
Lenfle, S. and Loch, C. (2017) Has Megaproject Management Lost Its Way?: Lessons from History. In B. Flyvbjerg (ed.) *The Oxford Handbook of Megaproject Management*, 21-38. Oxford University Press, Oxford, England.
Logan, D. C. (2009) Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry. *Journal of Experimental Botany*, Vol. 60, No. 3, pp. 712–714.
Mačák, K. (2017) From Cyber Norms to Cyber Rules: Re-engaging States as Law-makers. *Leiden Journal of International Law*, Vol. 30, No. 4, pp 877 – 899.
McGrath, R.G. and MacMillan, I.C. (2000) *The Entrepreneurial Mindset: Strategies for Continuously Creating Opportunity in an Age of Uncertainty,* The Fellows of Harvard College, Cambridge, MA.
Microsoft (2016) *From articulation to implementation: Enabling progress on cybersecurity norms*, White Paper, Microsoft, Redmond, Washington.
Mills, A. J., Durepos, G., and Wiebe, E. (Eds.) (2010) *Encyclopedia of case study research*. Sage Publications, Inc. Thousand Oaks, CA.
MIT Technology Review Insights (2021) Better cybersecurity means finding the "unknown unknowns." MIT, Cambridge, MA.
Ramesh, R.V. and Browning, T.R. (2014) A conceptual framework for tackling knowable unknown unknowns in project management. *Journal of Operations Management* Vol. 37 (4) 190-204.
Rumsfeld, D. (2011) *Known and Unknown: A Memoir*. Sentinel, Penguin Publishing Group, USA
Salet, W. (2021) Public Norms in Practices of Transitional Planning: The Case of Energy Transition in The Netherlands. *Sustainability* Vol.13, 4454. https:// doi.org/10.3390/su13084454
Statista, U.S. Companies and Cyber Crime, (2021). https://www.statista.com/study/12881/smb-and-cyber-crime-in-theunited-states-statista-dossier/.
Straub, J. (2021) Defining, evaluating, preparing for and responding to a cyber Pear Harbor. *Technology in Society*, Vol. 65, 101599.
Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (2017). Michael N. Schmitt (ed.) 2nd ed. Prepared by the International Group of Experts at the Invitation of the NATO Cooperative Cyber Defence Centre of Excellence (CCD COE), Cambridge University Press.
Temple-Raston, D. (2016) A Worst Nightmare' Cyberattack: The Untold Story Of The SolarWinds Hack. All Things Considered, National Public Radio (NPR). https://www.npr.org/2021/04/16/985439655/a-worst-nightmare-cyberattack-the-untold-story-of-the-solarwinds-hack.
U.S. Congress (2021) *Cyber threats in the pipeline: lessons from the federal response to the colonial pipeline ransomware attack*. 117th Congress (2021-2022).
U.S. Department of Justice (2022, Mar 24) Four Russian Government Employees Charged in Two Historical Hacking Campaigns Targeting Critical Infrastructure Worldwide. DOJ, Office of Public Affairs, Washington, D.C. https://www.justice.gov/opa/pr/four-russian-government-employees-charged-two-historical-hacking-campaigns-targeting-critical.

U.S. Senate (2001) Hearing Before the Select Committee on Intelligence of the United States Senate, Hack of U.S. Networks by a Foreign Adversary. 117th Congress, First Session. Tuesday, February 23, 2001.

U.S. Senate Historical Office (1946) Report of the Joint Committee on the Investigation of the Pearl Harbor Attack. (The Pearl Harbor Committee). U.S. Senate, Washington D.C.

United States Cyber Economic Espionage Accountability Act (CEEAA) Summary: H.R.2281 — 113th Congress (2013-2014).

Watkins, M.D., and Bazerman, M.H. (2003) Predictable surprises: the disasters you should have seen coming. *Harvard Business Review*, Vol. 81, No. 3, pp. 72-80, 140.

Weick, KE, and Sutcliffe KM. (2015) *Managing the Unexpected: Sustained Performance in a Complex World* (3rd ed.) John Wiley & Sons, Hoboken, NJ.

Weick, K.E., Sutcliffe. K.M. and Obstfeld, D. (2005) Organizing and the process of sensemaking. *Organization Science* Vol. 16 (4) pp. 409-421. Institute for Operations Research and the Management Sciences.

Complicity in Unlawful Offensive Cyber Operations Under International Law on State Responsibility

Samuli Haataja
Griffith Law School, Griffith University, Gold Coast, Australia

s.haataja@griffith.edu.au

Abstract: States are increasingly engaging in cybersecurity cooperation activities and providing support to other states in offensive cyber operations. While international cooperation is generally encouraged and many cybersecurity cooperation activities are lawful, there is also a risk of being complicit in the internationally wrongful acts of other states. This paper examines the risk of complicity in offensive cyber operations under international law on aiding or assisting. It argues that, while international law in this context applies to cyber operations by states, existing uncertainties and limitations around the key components of the law on aiding or assisting are compounded by competing interpretations about how international law generally applies to state conduct in cyberspace. The paper consists of four sections. Following the introduction in section one, section two outlines some of the ways in which states are cooperating in relation to cybersecurity and offensive cyber operations. Section three examines the key elements of international law on aiding or assisting as contained within article 16 of the International Law Commission's Articles on the Responsibility of States for Internationally Wrongful Acts, and the extent to which these apply or are problematised in relation cyber operations. It demonstrates that article 16 adopts a broad approach to what constitutes 'aiding or assisting' and this captures various types of activities in support of cyber operations provided the aid or assistance contributes significantly to a wrongful act of another state, the accomplice state has knowledge of the factual circumstances and the illegality of the act by the principal state, and where the accomplice state and principal state are bound by the same legal obligation. Section four concludes by outlining the limits of cooperation in the cyber context and how states can mitigate the risk of complicity in violations of international law.

Keywords: aid and assistance, cyber operations, international law, state responsibility

1. Introduction

This paper considers aiding or assisting in cyber operations under international law on state responsibility. It examines the risk of state complicity in violations of international law arising from cybersecurity cooperation. It argues that, while international law on aiding or assisting applies to cyber operations, existing uncertainties and limitations around the key elements of the law are compounded by competing interpretations about how international law generally applies to state conduct in cyberspace.

This paper consists of four sections. Section two outlines the various ways in which states are cooperating in relation to cybersecurity and cyber operations. Section three examines how international law on aiding or assisting, as contained in article 16 of the International Law Commission's (ILC) Articles on the Responsibility of States for Internationally Wrongful Acts (ARSIWA) (ILC 2001), applies to cyber operations. It demonstrates that article 16 covers a broad range of conduct and captures any form of aiding or assisting, whether this involves sharing mere lines of code, information about vulnerabilities, or malware provided there is a causal connection between the aid or assistance and the internationally wrongful act to which it contributes. However, the requirement under article 16 that the aiding or assisting state has knowledge of the circumstances of the wrongful conduct that they are contributing to is problematised by competing interpretations of how international law applies in cyberspace. Finally, the double obligation requirement within article 16 has potential to limit the application of the law in relation to regional cybersecurity treaties, and where states are persistent objectors to rules of customary international law in the cyber context. Section four provides a conclusion.

2. Cybersecurity and International Cooperation

States are increasingly cooperating in relation to cybersecurity activities and providing support to others engaging in cyber operations. For example, the recent AUKUS partnership between Australia, the United Kingdom (UK) and the United States (US) includes joint efforts to 'strengthen[] cyber capabilities' (White House 2022a) and Quad leaders have in 2022 agreed to 'strengthen information-sharing' among their Computer Emergency Response Teams (CERTs) (White House 2022b). The US also revealed it had conducted cyber operations in support of Ukraine in 2022 during its war with Russia. According to Paul Nakasone, this involved 'a series of operations across the full spectrum; offensive, defensive, [and] information operations' (Martin 2022). In the same context, it was noted that the US has been involved in sharing intelligence with its allies to undermine their adversaries' abilities to use their cyber capabilities (Martin 2022).

A number of states also discuss cooperation in their national positions on how international law applies to cyber operations. Canada, for example, in outlining its position on countermeasures, maintains that a state can provide assistance to another state requesting it where it 'does not possess all the technical … expertise to respond to internationally wrongful cyber acts' (Government of Canada 2022). Israel, in discussing the status of the due diligence principle, maintains that the nature of cyberspace 'incentivize[s] cooperation between States on a voluntary basis' and that CERTs regularly exchange information and cooperate to mitigate cybersecurity incidents (Schöndorf 2021, p. 404). As these examples and statements illustrate, states are cooperating in relation to different cyber activities and engaging in cyber operations in support of other states.

Cooperation around cybersecurity can involve a range of technical activities conducted for different purposes, many of which are lawful.[1] For example, states can share information about unknown vulnerabilities in software or hardware for the purpose of improving each others' cybersecurity. However, states can also share tools or techniques that exploit these vulnerabilities, and other malware utilising a range of exploits in order to obtain unauthorised access or disrupt the operation of adversary systems. Where these activities cause effects in the networks of another state, they can constitute violations of the international legal obligations of the state using these capabilities. However, there is also risk that the state involved in sharing these capabilities or providing technical support becomes complicit in violations of international law.

3. International Law on Aiding or Assisting

Under international law on state responsibility, a state (the accomplice state) can be responsible for aiding or assisting another state (the principal state) in an internationally wrongful act. Article 16 of the ARSIWA provides that:

> A State which aids or assists another State in the commission of an internationally wrongful act by the latter is internationally responsible for doing so if: (a) that State does so with knowledge of the circumstances of the internationally wrongful act; and (b) the act would be internationally wrongful if committed by that State. (ILC 2001)

This is echoed in the *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Tallinn Manual) where rule 18(a) provides that, in relation to cyber operations, a state is responsible for:

> its aid or assistance to another State in the commission of an internationally wrongful act when the State provides the aid or assistance knowing of the circumstances of the internationally wrongful act and the act would be internationally wrongful if committed by it[.] (Schmitt 2017, p. 100)

Despite some debate about whether article 16 is a primary or secondary rule of international law (Jackson 2015, p. 149), the International Court of Justice has stated that it reflects customary international law on aiding or assisting (ICJ 2007, p. 217). Therefore, while cooperation among states is generally encouraged, article 16 provides the limits of acceptable conduct when doing so (Jackson 2015, p. 158). In other words, it defines the boundaries between lawful cooperation among states and complicity in unlawful conduct that in itself violates international law.

Article 16 consists of four components. First, it requires conduct by the accomplice state that constitutes 'aid or assistance' to the principal state. Second, there must be a causal connection or nexus between the aid or assistance provided by the accomplice state and the wrongful conduct by the principal state. Third, the accomplice state needs to have 'knowledge of the circumstances of the internationally wrongful act' by the principal state. Finally, under the 'double obligation' requirement in article 16(b), the accomplice state must be bound by the same legal obligation that is being violated by the principal state. Each of these will be examined in turn, including how they apply and the extent to which they are problematised in the cyber context.

[1] Passive cyber defense measures, such as the use of firewalls, applying patches to software and other measures that do not cause effects outside a state's own territory, generally do not infringe on other states' rights under international law (Lahmann 2020, p. 125).

3.1 'Aiding or Assisting'

Article 16 adopts a broad approach to activities and conduct that can constitute 'aiding or assisting.' While the ILC did not define aiding or assisting, it provided a number of examples of conduct that can amount to such including providing essential facilities, financial support, closing international waterways, facilitating abduction of persons, or assisting in the destruction of property (Crawford 2002, p. 148). Similarly, a survey of state practice in this context provides examples such as the provision of military support, assistance in unlawful uses of force, foreign aid and economic cooperation, shelter and safe haven for unlawful activities, legal assistance, and extradition (Aust 2011, p. 108). As these examples illustrate, the conduct that can constitute aiding or assisting under article 16 is broad. Others adopt a similar position (see Jackson 2015, p. 154; Crawford 2013, p. 402). Aust argues that it is not possible to provide definitive list of what constitutes 'aid or assistance' or abstract criteria to determine it, and that what constitutes 'aid or assistance' is to be determined on a case-by-case basis (Aust 2011, pp. 210, 230). Similarly, the Tallinn Manual experts maintained that states can be responsible for aiding or assisting whether the conduct is 'cyber or non-cyber in nature' (Schmitt 2017, p. 100). Nonetheless, aiding or assisting generally requires positive acts by the accomplice state meaning omissions are excluded (Crawford 2013, pp. 403-5), as are situations where states simply provide 'advice, encouragement or incitement' to another state (Crawford 1999, p. 48). Accordingly, article 16 adopts a broad approach under which any kind of aid or assistance in cybersecurity cooperation activities can potentially fall within its ambit.

3.2 Causal Connection

There must, however, be a causal connection or nexus between the aid or assistance by the accomplice state, and the wrongful act by the principal state. According to the ILC, the accomplice state will only be responsible 'to the extent that its own conduct has caused or contributed to the internationally wrongful act' (Crawford 2002, p. 148).

As to how to determine whether the causal connection is sufficient, the ILC maintained that the aid or assistance does not need to be 'essential to the performance of the wrongful act' provided it 'contributed significantly' to the principal state's wrongful act (Crawford 2002, p. 149). Jackson maintains this requires 'material facilitation' of the wrongful act, as it 'catches conduct with a sufficient link to another state's wrongdoing while excluding the incidental relationships that arise from virtually every state interaction' (Jackson 2015, p. 158). But the aid or assistance cannot be too remote or indirectly related to the wrongful act (Nolte & Aust 2009, p. 10.) Similarly, a 'but for' test is not appropriate because if the aid or assistance is essential to the performance of the wrongful act, then the accomplice state is more likely to the main author of the wrongful act or jointly responsible (ILC 2002, art. 47).

In the cyber context, the Tallinn Manual experts adopted a similar position, noting that 'irrespective of the form it takes, [the aid or assistance] must materially and causally contribute to the act' (Schmitt 2017, p. 102). They maintained that aiding or assisting is a threshold above involvement that is 'merely tangential' to the wrongful act, but below the threshold of 'jointly conducting' it (Schmitt 2017, p. 102). As an example of when the aid or assistance crosses the latter threshold, the Tallinn Manual described a situation where 'a State provides uniquely effective and indispensable decryption capabilities to another State to enable the latter to conduct a harmful cyber operation, and without which the operation could not be mounted' (Schmitt 2017, p. 102). Boutin similarly argues that where the aid or assistance is 'critical to the main cyber operation', such as where it involves 'an essential facility or unique technical expertise', then that could give rise to joint responsibility and rise above the threshold of aiding or assisting under article 16 (Boutin 2019, p. 200).

Accordingly, a broad range of cybersecurity cooperation activities, such as support in conducting cyber operations that constitute internationally wrongful acts, can constitute aiding or assisting. This can involve sharing information about unknown vulnerabilities or techniques on how to exploit vulnerabilities, or provision of malware designed to cause particular effects. In each instance, even where the aid or assistance is seemingly minor from a technical perspective (such information about an unknown vulnerability), provided that it contributes significantly to the commission of an internationally wrongful act by the principal state, then the activity can constitute 'aid or assistance' under article 16. This, however, can be difficult to determine in the cyber context. For example, in some circumstances information about a single unknown vulnerability may be an important factor in the development of malware used in a cyber operation against another state and contributes significantly to the wrongful act for which it is used. In other circumstances, such as where the cyber operation involves malware exploiting multiple unknown vulnerabilities or involves a range of different tools and techniques, then sharing information about how to exploit a single unknown vulnerability could only be an

incidental factor that is too indirect to the wrongful act that the malware is used for. Therefore, combined with the lack of precision in the threshold of when aiding or assisting contributes significantly to a wrongful act, establishing or demonstrating that cybersecurity cooperation activities have met this threshold can be difficult.

3.3 'Knowledge of the Circumstances'

The third component is the requirement under 16(a) that the accomplice state has 'knowledge of the circumstances of the internationally wrongful act' by the principal state. This includes actual knowledge but does not include constructive knowledge (Crawford 2013, p. 406). Some also maintain that this includes where there is 'wilful blindness' on behalf of the accomplice state (Jackson 2015; Moynihan 2016; Milanovic 2021). According to the commentary to the ILC articles, the accomplice state must be 'aware of the circumstances making the conduct of the assisted State internationally wrongful' (Crawford 2002, p. 149). This can be interpreted to mean that the accomplice state must have knowledge of both the factual circumstances and illegality of the conduct by the principal state (Moynihan 2016, pp. 11-1; Lanovoym 2016, p. 100).

However, there is a degree of uncertainty about whether 'knowledge of the circumstances of the internationally wrongful act' also requires intention on behalf of the accomplice state to aid or assist the principal state in the wrongful act, or whether knowledge is sufficient. This uncertainty arises because the ILC commentary provides that the aid or assistance by the accomplice state 'must be given with a view to facilitating the commission of the wrongful act' by the principal, and the commentary notes that accomplice state will not be responsible unless it 'intended, by the aid or assistance given, to facilitate the occurrence of the wrongful conduct' (Crawford 2002, p. 149). Further, the reference to intention in the commentary to article 16 is inconsistent with the approach to 'intent' that the ARSIWA generally adopt (Crawford 2002, p. 84; Nolte & Aust 2009, p. 13).

This uncertainty has given rise to some debate about whether or not article 16 requires a degree of intent by the accomplice state, or whether knowledge is sufficient (Jackson 2015, pp. 160-1). However, among those who do argue that intention is required in addition to knowledge (Crawford 2013, pp. 407-8; Aust 2011, pp 237-9; Moynihan 2016, pp. 18-9; Milanovic 2021, p. 1321), there is a tendency to adopt a broad or flexible approach to determining this. For example, Nolte and Aust emphasise the importance of intention but maintain that a flexible approach should be taken to prevent states from hiding behind the intent requirement, and that in some cases 'a lack of intent can be offset by sufficient knowledge' (Nolte & Aust 2009, p. 15). Crawford also maintains that intent can be imputed where the aid or assistance is provided 'with certain or near-certain knowledge as to the outcome' (Crawford 2013, p. 408). Similarly, Moynihan maintains that intention is required but concludes that this can be satisfied where the accomplice state has 'knowledge or virtual certainty that the recipient state will use the assistance unlawfully' Moynihan 2016, p. 20). Milanovic adopts a similar approach under which the accomplice state has 'indirect intent' where it chooses to provide aid or assistance in circumstances that it is 'practically certain' it would facilitate the commission of a wrongful act by the principal state (Milanovic 2021, p. 1321). Accordingly, even where article 16 is interpreted to require intention in addition to knowledge, most adopt a broad approach to determining this which highlights the importance of the accomplice state having knowledge in terms of a degree of certainty about factual circumstances and the illegality of the conduct in question.

In the cyber context, article 16's knowledge requirement is complicated by the competing interpretations about how international law applies to state activities as this affects whether the accomplice state has knowledge of the illegality of the conduct for which they are providing aid or assistance. For example, while most states agree that sovereignty operates as a rule of international law that can be violated, there continues to be uncertainty about the threshold at which a violation of sovereignty occurs (see Heller 2021, pp. 1458-64). Consider, for instance, where a state provides aid or assistance in the form of cyber capabilities that help the principal state to obtain unauthorised access to computers located within another state with the aim of taking down a botnet being operated from those systems. In this situation, the accomplice state may have knowledge of the relevant factual circumstances, but it may not have knowledge about the illegality of the conduct as it is unclear whether the principal state's conduct would violate sovereignty and be unlawful given the different positions that states have adopted to determining this.

3.4 The Double Obligation Requirement

The final component is the so called 'double obligation' requirement under article 16(b). This is based on the *pacta tertiis* rule (that treaties only create obligations to the parties to it, and not third parties) and requires that

the accomplice state and the principal state are bound by the same legal obligation that is being violated and that the accomplice state is aiding or assisting in (Crawford 1999, pp. 50-1).

The effect of the double obligation requirement is that it limits the scope of article 16. For example, in relation legal obligations arising from bilateral or regional treaties, states not party to these treaties may aid or assist in the breach of them without being responsible under article 16. Generally, in relation to obligations arising from customary international law, all states will be bound by the same obligation and satisfy the requirement in article 16(b). However, states that are persistent objectors and not bound by the general rule fall outside the scope of article 16(b) where they aid or assist in the wrongful act of another state (Jackson 2015, pp. 162, 167).

In the cyber context, an example of how the double obligation requirement may limit the application of article 16 arises from recent changes to European Union (EU) export controls over 'dual use' cyber capabilities (EU 2021). Under the new EU regulation, obligations are imposed on EU member states to limit to the export of dual use 'cyber-surveillance items' where they are or may be intended for use in connection with 'serious violations of human rights and international humanitarian law' (EU 2021, art. 5). In circumstances where a non-EU state aids or assists an EU member state in violations of the law, such as by contributing to the development of cyber-surveillance items with knowledge they will be exported, this would not be captured by article 16 given the requirement that both states are subject to the same legal obligation. While the EU regulation is not a regional treaty, it illustrates how there may be circumstances where a regional cybersecurity treaty only imposes obligations on certain states, and how the aid or assistance provided by those not party to the treaty would fall outside the scope of article 16.

Another example arises from persistent objectors to rules of customary international law. In 2020 NATO released its 'Allied Joint Doctrine for Cyberspace Operations' (AJP-3.20) in which its members expressed their agreement to a number of questions of how international law applies to cyber operations (NATO 2020). NATO's AJP-3.20 provides, for example, that cyber operations below the use of force threshold 'may nevertheless constitute a violation of international law as a breach of sovereignty' (NATO 2020, p. 20, fn. 26), and many states have expressed support for a similar position (see UN 2021). The UK, however, has since 2018 maintained that this is not the case, and this was reiterated in NATO's AJP-3.20 where the UK made a reservation to this effect (NATO 2020, p. v). Given that states have begun to carve out new understandings of how international law applies to their cyber activities (see Akande, Coco & Dias 2022, p. 24) and documents such as this can be regarded as evidencing state practice and *opinio juris* for this purpose, the UK's position can be interpreted as an effort to persistently object to the application of the rule of territorial sovereignty to cyber operations below the threshold of the non-intervention and non-use of force principles. In this context, and assuming a cyber specific rule on sovereignty develops into customary international law and that the UK would not be bound by this rule, then the application of article 16 is limited in relation to those persistent objectors. For example, where the UK were to aid and assist another state through the provision of cyber capabilities which contributed significantly to a cyber operation in violation of another state's sovereignty, the UK's conduct could fall outside the scope of article 16 given the double obligation requirement in article 16(b).

3.5 Article 16 and the Limits of Cooperation in the Cyber Context

Accordingly, article 16 adopts a broad approach to aiding or assisting which is neutral to the technologies used in this context. In relation to cybersecurity cooperation activities, the effect of this is that even where the provision of aid or assistance involves sharing lines of malicious code or information about unknown vulnerabilities, this can fall within the scope of article 16. But the key question is whether the aid or assistance by the accomplice state 'contributed significantly' to an internationally wrongful act by the principal state. This may be difficult to establish in the cyber context but will be considered on a spectrum. The aid or assistance must be more than 'merely tangential' participation (Schmitt 2017, p. 102) or an 'incidental factor' in the commission of the wrongful act (Crawford 2002, p. 151). But it must be less than 'essential to the performance of the wrongful act' or where the states are jointly conducting the wrongful act (Crawford 2002, p. 149; Aust 2011 pp. 212-3).

As to the requirement that the accomplice state has 'knowledge of the circumstances of the internationally wrongful act', the uncertainty around what extent of knowledge is required by the accomplice state is compounded in the cyber context. Essentially, while an accomplice state may have knowledge of the factual circumstances, it may be difficult to have knowledge of the illegality of the conduct by the principal state given that states have not yet reached consensus on how relevant primary rules of international law apply in the cyber context and the thresholds at which cyber operations will violate them (see UN 2021). Finally, the double

obligation requirement has potential to limit the scope of article 16 particularly in relation to obligations arising from regional treaties, and where states are persistent objectors to cyber specific rules of customary international law. This also demonstrates how debates about the status and scope of rules of international law in the cyber context, such as sovereignty, can have flow-on effects into other areas of law, such as aiding or assisting.

To mitigate the risk of complicity in internationally wrongful acts arising from cybersecurity cooperation activities, states should consider the circumstances surrounding the aid or assistance they are providing, and how the state they are aiding or assisting interprets international law to apply to cyber operations. Where there are concerns around the risk of complicity, the accomplice state can mitigate this by, for example, imposing conditions on the aid or assistance they provide (Moynihan 2016, pp. 38-44; Jackson 2015, p. 161). Further, given how the uncertainties in the law on aiding or assisting are compounded by debates about the status and scope of other rules of international law in the cyber context, it remains imperative for states to continue developing clarity about how they consider international law to apply in the cyber context, particularly in relation to rules such as sovereignty which can delineate the boundaries between lawful cyber espionage and unlawful cyber operations below the use of force threshold.[2]

4. Conclusion

This paper examined aiding or assisting in cyber operations under international law on state responsibility. It demonstrated how article 16 adopts a broad approach to what constitutes 'aiding or assisting' and how this captures various types of activities in support of cyber operations. While cybersecurity cooperation activities are generally lawful, there is a risk of complicity where the aid or assistance contributes significantly to a wrongful act of another state, the accomplice state has knowledge of the factual circumstances and the illegality of the act by the principal state, and where the accomplice state and principal state are bound by the same legal obligation. States can mitigate these risks by imposing conditions on the aid or assistance they provide, and by developing further clarity in their national positions on how international law, including the law on aiding or assisting, applies in the cyber context.

References

Akande, D, Coco, A, & Dias, TS 2022, 'Drawing the Cyber Baseline: The Applicability of Existing International Law to the Governance of Information and Communication Technologies', *International Law Studies*, vol. 99, pp 4-36.

Aust, HP 2011, *Complicity and the Law of State Responsibility*, Cambridge University Press, Cambridge.

Boutin, B 2019, 'Shared Responsibility for Cyber Operations', *AJIL Unbound*, vol. 113, pp 197-201, DOI:<https://doi.org/10.1017/aju.2019.31>.

Buchan, R 2019, *Cyber Espionage and International Law*, Hart, Oxford.

Crawford, J 1999, *Second Report on State Responsibility*, UN Doc A/CN.4/498.

Crawford, J 2002, *The International Law Commission's Articles on State Responsibility: Introduction, Text and Commentaries*, Cambridge University Press, Cambridge.

Crawford, J 2013, *State Responsibility: The General Part*, Cambridge University Press, Cambridge.

Delerue, F 2020, *Cyber Operations and International Law*, Cambridge University Press, Cambridge.

European Union (EU) 2021, Regulation (EU) 2021/821 of the European Parliament and of the Council of 20 May 2021 setting up a Union regime for the control of exports, brokering, technical assistance, transit and transfer of dual-use items (recast), 2021 OJ (L 206) 1.

Government of Canada 2022, *International law applicable in cyberspace*, 4 April, viewed 30 January 2023, <https://www.international.gc.ca/world-monde/issues_development-enjeux_developpement/peace_security-paix_securite/cyberspace_law-cyberespace_droit.aspx?lang=eng>.

Heller, KJ 2021, 'In Defense of Pure Sovereignty in Cyberspace', *International Law Studies*, vol. 97, pp 1432-1499.

International Court of Justice (ICJ) 2007, *Case Concerning Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia-Herzegovina v Yugoslavia)* (Judgment) [2007] ICJ Rep 43.

International Law Commission (ILC) 2001, *Responsibility of States for Internationally Wrongful Acts*, UNGA Res 56/83, 12 December, UN Doc A/RES/56/83.

Jackson, M 2015, *Complicity in International Law*, Oxford University Press, Oxford.

[2] Peacetime espionage is not directly addressed by international law and cyber espionage activities, depending on their effects, are generally not considered to violate international law (Schmitt 2017, p. 25). For those who argue that cyber espionage does violate international law on sovereignty, see, for example, Buchan (2019, p. 54) and Delerue (2020, p. 214).

Lahmann, H 2020, *Unilateral Remedies to Cyber Operations*, Cambridge University Press, Cambridge.
Lanovoym, V 2016, *Complicity and its Limits in the Law of International Responsibility*, Bloomsbury, Oxford.
Martin, A 2022, 'US military hackers conducting offensive operations in support of Ukraine, says head of Cyber Command', *Sky News*, 1 June 2022, viewed 30 January 2023, <https://news.sky.com/story/us-military-hackers-conducting-offensive-operations-in-support-of-ukraine-says-head-of-cyber-command-12625139>.
Milanovic, M 2021, 'Intelligence sharing in multinational military operations and complicity under international law', *International Law Studies*, vol. 97, pp 1269-1403.
Moynihan, H 2016, *Aiding and Assisting: Challenges in Armed Conflict and Counterterrorism*, Chatham House Research Paper, viewed 30 January 2023, <https://www.chathamhouse.org/sites/default/files/publications/research/2016-11-11-aiding-assisting-challenges-armed-conflict-moynihan.pdf>.
North Atlantic Treaty Organisation (NATO) 2020, *Allied Joint Doctrine for Cyberspace Operations, Allied Joint Publication-3.20*, NATO Standardisation Office, viewed 30 January 2023, <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/899678/doctrine_nato_cyberspace_operations_ajp_3_20_1_.pdf>.
Nolte, G & Aust, HP 2009, 'Equivocal Helpers—Complicit States, Mixed Messages and International Law' *International & Comparative Law Quarterly*, vol. 58, no. 1, pp 1-30, DOI:<https://doi.org/10.1017/S0020589308000821>.
Schmitt, M (ed.) 2017, *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*, 2nd edn, Cambridge University Press, Cambridge.
Schöndorf, R 2021 'Israel's Perspective on Key Legal and Practical Issues Concerning the Application International Law to Cyber Operations', *International Law Studies*, vol. 97, pp 395-406.
United Nations (UN) 2021, *Official compendium of voluntary national contributions on the subject of how international law applies to the use of information and communications technologies by States submitted by participating governmental experts in the Group of Governmental Experts on Advancing Responsible State Behaviour in Cyberspace in the Context of International Security established pursuant to General Assembly resolution 73/266*, 13 July, UN Doc A/76/136*.
White House 2022a, *Fact Sheet: Implementation of the Australia – United Kingdom – United States Partnership (AUKUS)*, Press release, 5 April, viewed 30 January 2023, <https://www.whitehouse.gov/briefing-room/statements-releases/2022/04/05/fact-sheet-implementation-of-the-australia-united-kingdom-united-states-partnership-aukus/>.
White House 2022b, *Fact Sheet: Quad Leaders' Tokyo Summit 2022*, 23 May, viewed 30 January 2023, <https://www.whitehouse.gov/briefing-room/statements-releases/2022/05/23/fact-sheet-quad-leaders-tokyo-summit-2022/>.

Semiotics of Strategic Narratives of "Antichrist" in Russia's War in Ukraine

Michael Bennett Hotchkiss
Independent Researcher, Griswold Connecticut, USA

mbhotchkiss@gmail.com

Abstract: The Russian war in Ukraine which began on February 24, 2022 coincides with the ongoing schism of the Moscow Patriarchate of the Russian Orthodox Church from the Ecumenical Patriarchate of Constantinople, to which the Orthodox Church of Ukraine affiliates. In this setting, spiritual and secular leaders in Russia and Ukraine have mutually utilized narratives of the Antichrist and Satan to explain Russian attacks on Ukraine, imbuing a "spiritual" dimension to the strategic communications in the conflict. This paper applies a semiotic approach to analysing the Antichrist and satanic myths at play in the context of ideological "strategic narratives", and the conflict of meanings which emerges from these competing narratives. In Russia, these ideomyths have long been utilized as an ideological tool which places the nation metaphysically in perennial militant opposition to the West. However, Ukraine which is striving to leave the Russian orbit and join the West has reciprocally framed Russia and its leader in similar concepts. In conclusion, this paper argues that there is opportunity for the creation of political messaging which can frame the conflict in spiritual and moral terms that can resonate with both the West and Russian thinkers.

Keywords: Antichrist, Ideomyth, Information Geopolitics, Semiotics, Strategic Narratives

1. Introduction

This review article analyses competing wartime strategic narratives of "Antichrist" and "Satan" used by senior public figures in contemporary Ukraine and Russia which were openly reported in news media during the year of 2022.

The article offers an innovative theoretical perspective by integrating these observations from 2022 with a semiotics-focused study of the Antichrist as a key "ideomyth" of post-Soviet Russian culture (Dolinska-Rydzek 2021), and also with a semiotics-focused approach to the analysis of "strategic conspiracy narratives" (Madisson and Ventsel 2020).

The semiotic approach to analysis of strategic conspiracy narratives in an information warfare context (such as these public statements about the Antichrist may be theoretically considered) takes a primarily cognitive security focus (Madisson & Ventsel 2020).

According to Lotman (2005), a semiosphere can be defined as the smallest functioning unit of semiosis, comprised of hierarchically organized sub-structures (human personalities, memory images, texts, etc.) that comprise the entire "semiotic space" of a given culture. It exists in a continuum of other semiospheres representing other cultures.

The author posits that since the fall of the Soviet Union, the cultural semiospheres of Ukraine and Russia are increasingly ideologically divergent. This is evidenced in the wartime meaning of the Antichrist ideomyth concepts as projected by the countries' respective churches and secular security leaders.

The article first presents the context of the Antichrist ideomyth in 2022 as reflected in worldwide Google Trends results, which demonstrates a direct linkage to Russia's war of aggression in Ukraine, and the respective leaders of Ukraine and Russia. Next, the competing narratives related to the Antichrist which have been offered by Ukrainian and Russian religious and secular security figures during the war are presented.

The literature on the Antichrist ideomyth in post-Soviet Russia is then discussed. This introduces key ideological figures involved in the ongoing production of the ideomyth. The literature on strategic conspiracy narratives is next considered, which relates the 2022 statements to the Russian field of "information geopolitics" which relies on the Russian Orthodox Church as a tool of sharp power projection.

Finally, the outlines of a "pro-Western" strategic conspiracy narrative are proposed by the author which would follow the lead of the Ukrainians and leverage the deeply-ingrained global semiotics of the Antichrist ideomyth in order to symbolically promote a rules-based international order.

2. The Antichrist of 2022

In February 2022, worldwide interest in the topic "Antichrist" spiked to its highest level since 2020 on Google Trends (Figure 1). There was a clear association between the commencement of Russia's "Special Military Operation" and the virality of Antichrist as a search term. The associated results show related queries are most often for Vladimir Putin, President of Russia; and Volodymyr Zelenskyy, President of Ukraine. This indicates that internet users were most likely wondering if these men might be candidates for the identity of the end times Antichrist.

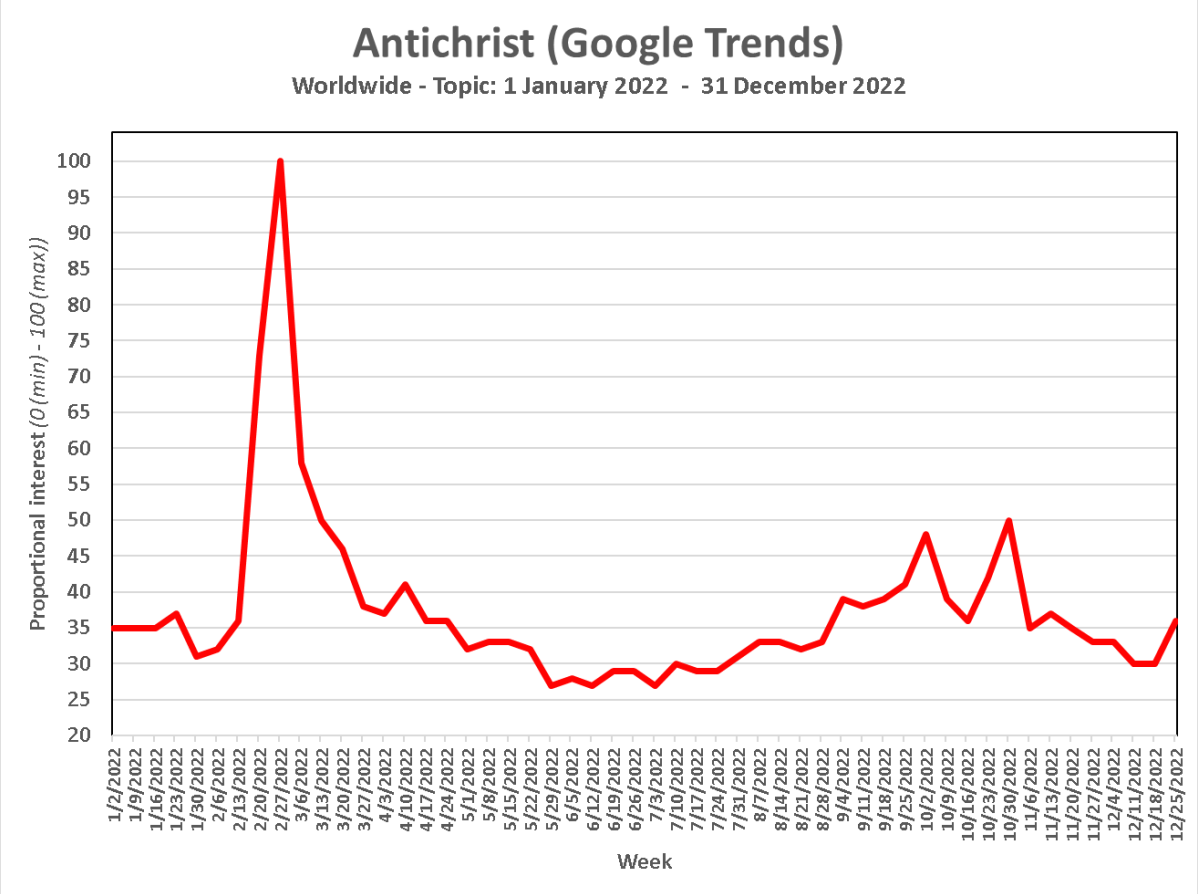


Figure 1: Google Trends data for "Antichrist" (Topic, Worldwide) - 2022

As indicated by the sudden shift in popular conspiratorial suspicion of Bill Gates as Antichrist during the coronavirus pandemic to that of Putin or Zelenskyy as Antichrist during the Special Military Operation, the Antichrist is an historical "floating signifier" (Almond 2020), whose identity changes depending on the contemporary cultural and political circumstances.

In nearly two millennia since the scribing of 1 John 2:18, which reads (KJV): "*Little children, it is the last time: and as ye have heard that Antichrist shall come, even now are there many Antichrists; whereby we know that it is the last time*", observers wondering if they are living in the end times have speculated on a seemingly-unending list of potential suspects. Such an enduring global fascination with the figure of the Antichrist suggests the term may have perennial plasticity for purposes of political warfare.

2.1 The Ukrainian position: Putin as the Antichrist of our current times

On September 6, 2014 in response to Russia's invasion of 'fraternal' Ukrainian territories, primate of the Ukrainian Orthodox Church (Kyiv Patriarchate), Patriarch Filaret (Denysenko) published a letter implying that Vladimir Putin faced "eternal damnation" and was "*a real new Cain – not by name, but by deeds*" … "*we know that two great sins, murder and lie are in close unity with each other, as the ruler and the inspiration of both is the devil. Therefore, anyone who indulges in these sins, according to the Savior, has the devil as his father.*" … "*His lie is misleading some people, and they think that in fact this ruler protects traditional spiritual and moral*

values from the ravages of globalization. But the fruit of his actions, which the Gospel calls us to evaluate, suggest otherwise." … "*Calls for this ruler and his minions have already been sounded many times publicly - to think again, to stop sowing evil and death, to repent. But it seems that he remains deaf to these appeals and only multiplies evil, because satan went into him, as into Judas Iscariot."* The Patriarch further compared the Russian leader to the Pharaoh, and the Ukrainians' plight to the freedom from enslavement of the Jews in ancient Egypt (Denysenko 2014).

In October 2018, this spiritual argument between the brotherly nations of Ukraine and Russia significantly escalated when the Ecumenical Patriarchate of Constantinople indicated its intention to grant autocephaly to a new Ukrainian Church. This resulted in the largest schism in modern Orthodox Christian history, when the Russian Orthodox Church (Moscow Patriarchate) broke off communion with the Ecumenical Patriarchate in response.

The new Orthodox Church of Ukraine (OCU) was established on December 15, 2018, and on January 5, 2019, Ecumenical Patriarch Bartholomew I signed the *tomos* granting the church official autocephaly.

On February 27, 2022 (the fourth day of the new war), and shortly after Google Trends interest in "Antichrist" seems to have peaked, Metropolitan Epiphanius (Dumenko), the first primate of the OCU, wrote a scathing letter about Putin, noting the condemnation of the war by other Orthodox Church leaders, and likening the Russian leader to both Antichrist and Hitler for his nuclear threats.

"*Today, the President of Russia ordered to increase the level of readiness of nuclear weapons. Isn't this a confirmation of the fact that the head of our state Volodymyr Zelenskyy and many others have said from the beginning: not only Ukraine is in danger, the whole world is in danger. The spirit of the Antichrist operates in the leader of Russia, the signs of which the Scriptures reveal to us: pride, devotion to evil, ruthlessness, false religiosity. This was Hitler during World War II. This is what Putin has become today.*" (Dumenko 2022)

Archbishop Yevstratiy Zoria, a spokesman of the OCU also told the BBC on February 27 in an interview that: "*Putin is really not messiah, but really Antichrist of our current time… He is Antichrist because everything what he does, everything what he does now is totally against Gospel, against God's law.*" (Jansezian 2022)

2.2 The Russian position: a unifying campaign to "desatanize" Ukraine

A week after the March 6, 2022 lead of Patriarch Kirill (Gundyaev) who then alleged that opposition to "gay parades" provided the theological basis for the war, the head of Chechnya, Ramzan Kadyrov seems to have invented the term "de-shaitanize" in relation to Ukraine (Ms.Detector.Media 2022, Orlova 2022).

In a June interview with the pro-Kremlin Serbian news outlet IN4S, which Montenegrin authorities alleged to be "*the propaganda branch of the Serbian Information and Security Agency (BIA)*", close associate of Kadyrov, Secretary of the Security Council of Chechnya, and commander of the Chechen Akhmat detachment, General Apti Alaudinov said: "*I call the Akhmat special forces the army of Isa, the army of Jesus. Because we, true Muslims, are waiting for the coming of Isa-Jesus more than anyone else. Because we - both Muslims and Christians - we are one, whole. And when we unite, on this day the devil and his army will cry.*" (Krasno 2022)

When announcing his annexation of Ukrainian territories on September 30, 2022, Vladimir Putin declared his opponents were followers of "*a religion on the contrary—outright Satanism,*" borrowing rhetoric normally associated with the Russian Orthodox Church and Russian anti-cult movement (Introvigne 1, 2022).

On October 25, 2022, the call for the "desataniziation of Ukraine" was first mentioned in news reports by Aleksey Pavlov, Deputy Secretary of the Security Council who also borrowed rhetoric from the anti-cultist movement (Introvigne 2, 2022).

On the same day, Patriarch Kirill (Gundyaev) called Putin "a fighter against the Antichrist" in his remarks at the 24th Congress of the World Russian People's Council. At this same conference, was the session "*ideological war against Satan*" which included speakers such as Alexander Dugin (MS.Detector.Media 2022).

Also on October 25, 2022, Head of Chechnya, Ramzan Kadyrov called for a union of Christians and Muslims, and said the Special Military Operation was a "Jihad" in defence of traditional values. He exhorted his forces to "*destroy these shaitans, wherever they are and no matter how well they hide.*" (Kadyrov 2022)

On Russia's "Unity Day" on November 4, 2022, security council member and former president Dmitry Medvedev said the war's goal was "*to stop the supreme ruler of hell, whatever name he uses: Satan, Lucifer or Iblis*." (Iblis is a Muslim equivalent name for Satan.) Medvedev's use of multireligious terms appealing to various Abrahamic

faiths in describing the adversary of Russia was a notably "ecumenical" use of the desatanization idea following Putin's announcement at the annexation event (Magnenou 2022).

It was noted by analysts that "*desatanization is practically the only way to unite everyone*" and was "*a desperate search for an ideology that will create a single platform for the Russians.*" Further it was noted that Medvedev's speech was written with "*language taken from "A Thousand and One Nights", with quotes from Aitmatov. A colourful orientalism appears there, and all this, on the one hand, may indicate that they are trying to attract the attention of Muslim communities...*" (Medvedev 2022)

Although Aleksey Pavlov's claims sparked controversy in the Jewish community for its portrayal of Hasidism as a "hypersect cult" he alleged was connected to Ukrainian satanism, Pavlov also argued that Ukraine was planning to ban values associated with traditional faiths like Christianity, Islam, and Judaism in his original argument. (Pavlov was fired on January 20, 2023 at the order of Vladimir Putin without explanation for what many suspect was a popular backlash over his antisemitic remarks (Halpern 2023).)

It seems clear that this sequence of narratives delivered by high level figures from Russian religious and secular security circles represent attempts to reinforce a "cultural code" related to the Antichrist which may be useful for the popular unification of various domestic religious denominations in Russia in opposition to an all-encompassing common enemy; and may be particularly effective in creating favourable cognitive conditions that more readily encourage Russian soldiers to fight (and die) in Ukraine.

3. Literature review

3.1 The Antichrist as a Russian ideomyth

From a semiotic perspective, myth may be considered an "ideologically driven narrative" or the "most elementary form of the narrative", governed by underlying rules of the culture which produces it (Lincoln 1999). Specifically, an ideomyth refers to a "*culture mirroring and culture-forming narrative, adjusting its objectives to the contexts in which it is employed.*" (Dolińska-Rydzek 2021)

Dolińska-Rydzek (2021) considers the Antichrist to be an ideomyth which is "*an essential part of Russian cultural imagery… a significant part of historiosophic conceptions such as Moscow the Third Rome, or the Holy Rus*", or the idea of "Katechon". It supports "*an ideological narrative, which simultaneously reflects and constitutes Russian culture*" which "*the Antichrist figure encapsulates the most important features of*", having "*undergone numerous historical and semantic transformations throughout the centuries*" to suit the contemporary political requirements. The Antichrist is a "*multidimensional character*" which has "*served as a constitutive other and the final enemy against whom Russian national identity has been constructed*" and has "*served as a myth consolidating the community in the face of crisis.*"

The post-Soviet ideological conception of Antichrist which dominates during the "Special Military Operation" seems primarily related to the work of "Russian Orthodox nationalists" who utilise eschatological and anti-Western themes in their writing. This list includes (but is not limited to) public figures such as Vardan Bagdasaryan, Alexander Dugin, Konstantin Dushenov, Tatyana Gracheva, Metropolitan Ioann Snychev (deceased 1995), and Alexander Prokhanov (Dolińska-Rydzek 2021).

While many of these figures were once politically fringe, their works have increasingly been legitimised in mainstream culture. Those like Bagdasaryan, Dugin, and Prokhanov are believed to support the development of national ideologies for the Kremlin used in the war (Faure 2022, Laruelle 2022, Pertsev 2022).

According to Dolińska-Rydzek (2021), for today's Russian Orthodox nationalists the Antichrist "*denotes not only the forthcoming global apocalypse and the decay of the West, but also the complexity of Russian-Western relations and Russian supremacy as the only genuinely Orthodox power. What is more, the Antichrist represents phenomena such as globalization, ecumenism, new technologies, and anti-Russian plots*" … "*it is a result of unceasing processes of re-interpreting and re-working of apocalyptic and eschatological narratives, and aligning them to existing reality.*"

As noted by Bagdasaryan and Resnyansky (2016), "*each epoch in the development of Russian historiosophy created its own Antichrist, who had little in common with the beast from The Revelation of John the Theologian.*"

In accordance with this ever-regenerating system of meanings around the ideomyth of Antichrist and its 'floating signifier' status, throughout history, the external "metaphysical enemy" of the Russian Katechon has taken many

forms: *"the Tatars, the Turks, freemasons, Napoleon, Hitler, and nowadays American agents, Ukrainian fascists, and the Kiev Junta*." (Engström 2014)

However, Dolińska-Rydzek (2021) notes that in addition to the popular idea that Russia is Katechon, there is also a deeply rooted and contrasting idea within Russia that the embodiment of the Antichrist would be born there.

Suspects for the identity of the Russian Antichrist have at times included Ivan the Terrible, Peter the Great, Joseph Stalin, and Mikhail Gorbachev (Bagdasaryan & Resnyansky 2016, Isupov 1995, Dolińska-Rydzek 2021).

3.2 Strategic narratives of Antichrist and semiotic conflict in information geopolitics

Despite the mutual significance of these popular-eschatological symbols in the Ukrainian and Russian semiospheres, the author posits that it is evident that the diverging strategic narratives considered in the article related to the identification of the Antichrist/Satan are artifacts of a "semiotic conflict" in political meaning produced by the respective wartime cultures.

According to semiotic theory, in the margins between semiospheres is where both communication and conflict can occur (Lotman 2005, Madisson & Ventsel 2020). This interaction between spheres can transform meanings of a sign such as the complex identity of the Antichrist may be considered in the context of an ideomyth (Dolińska-Rydzek 2021).

Information geopolitics is a term attributed to Russian information warfare theorist Igor Panarin (Maliukevičius 2007). This field is also relevant to the examination of the work of philosopher Alexander Dugin, who in addition to Panarin, has been considered one of the primary Russian thought leaders of information geopolitics. Despite basic differences, both Panarin and Dugin similarly base their theories around direct competition with the Western system and the expansion of Russia's informational and/or territorial sphere of influence (Darczewska 2014).

According to Russian information geopolitics theorist Andrei Manoilo, penetration into other states' information environments is accomplished only where those states operate a mutual common "code" (language, religion, historical experience, etc.), and if those codes do not match, information confrontation can result (Maliukevičius 2007).

The author strongly feels this corresponds with the theoretical idea of semiotic conflict between semiospheres, and the transmission of Antichrist conspiracies through Orthodox Church channels.

Whereas traditional geopolitical theory understands power as derived through the management of a physical political geography, proponents of information geopolitics instead posit that geopolitical power is derived through control of the information space using techniques such as knowledge management, information infrastructure control, and the influence of public opinion (Darczewska 2014, Maliukevičius 2007).

In Russia, these projects in relation to figures like Panarin and Dugin have been consistently focused on the neo-imperial "information expansion" of "Russian Civilization" (or the "Russian World"), often in association with the Russian Orthodox Church (Maliukevičius 2007).

Bagdasaryan & Resnyansky (2016) note the geopolitical dimensions of the Antichrist ideomyth in Russia again evolved after the fall of the USSR to encompass opposition to "*a monocentric model of the world*", and the popular equation of the Antichrist with the "Anglo-American civilization", NATO, and the global financial system. They provide several examples of eschatological conspiracy theories which circulated in Russia's Trinity-Sergius Lavra monastery and may have influenced the behaviour of churches abroad, such as in Cyprus, Greece, and Finland.

On this basis, the author asserts it is plausible to see how the ideomyth of Antichrist can fit as a tool of information geopolitics spread through the "Russian World" which may influence internal and external audiences.

As Dolińska-Rydzek (2021) notes based on an analysis of works of the semioticians of the Tartu-Moscow Semiotic School, one of the "essential aspects" of the Orthodox Christian culture of Russia emphasises a kind of "Manichean vision" where "*the world is viewed in a paradigm either/or: either holy or diabolical, either good or bad, either Christian or anti-Christian*" and "*in this world-view, the West represents evil, while Russia is perceived as an embodiment of positive values.*"

In their book Strategic Conspiracy Narratives: A Semiotic Approach, Madisson & Ventsel (2020) treat conspiracy theory within the context of strategic communications and identity creation. The researchers state that *"influencing activities focused on the cognitive level of the contemporary information war are usually based on the inclusion and exclusion logic of identity creation" … "The communication of conspiracy theorists is characterized by strongly polarised identity creation in which an antagonistic opposition of "us" and "them" serves as an important dominant."*

Russian nationalist Bagdasaryan (2016) posits that "*to disavow geopolitical opponents, it is necessary to create an associative series that leads to their identification with images perceived as an unambiguous evil. Such an evil in medieval semiotics was the figure of the Antichrist…. The modern world is built on a value consensus of condemnation of fascism and Nazism. The role that the Antichrist played for the medieval consciousness is played today by Adolf Hitler. Accordingly, in order to expose the enemy, it should be likened to Hitler.*"

The author suggests that as the ultimate contrast to "good", ideomyths of the Antichrist and Satan have clear value for polarising identity creation, supporting the "Manichean vision" of the Russian Orthodox worldview in the context of strategic conspiracy narratives.

3.3 Security implications outside Russia and Ukraine

In the divergence of the Ukrainian church and political system from the Russian one – as well as apparently in response to over 20 years of the development of aggressive, nationalist rhetoric in Russia, we see potential semiotic conflict arising in differing interpretations of codes of "Antichrist". The author feels this conflict in meanings has implications for information and cognitive security in Western societies.

The Antichrist ideomyth narratives may fuel international extremist political activities on the far right, where Russia has demonstrated a significant investment in such groups (Dolińska-Rydzek 2021).

To support such a view, from the start of Kadyrov's use of desatanization rhetoric, there was also an intersection with the defence of traditional values and conspiratorial statements about gay parades. Further, the massive legitimisation of an irrational campaign against Satan seems likely to provide fodder for Christian nationalist groups, or supporters of fringe movements like QAnon (Lecaque 2022, Magnenou 2022).

The ideology of the "Russian World" driving the SMO which has been observed as a heresy by Orthodox leaders outside of the Russian Orthodox Church orbit is seductive to many in the "*Orthodox Church and has even been taken up by the Far Right and Catholic and Protestant fundamentalists.*" (Gallaher & Kalaitzidis 2022)

Therefore, Western societies may be interested in defending against and disrupting such narratives from an informational-cognitive security perspective, as these ideas seem likely to spread beyond Orthodox Christian semiospheres and may influence various extremist secular and religious groups.

4. Discussion: Creating a pro-Western strategic narrative of "desatanization"

On November 22, 2022, Ukrainian secret services raided the Kyiv-Pechersk Lavra monastery, which had been leased by the Ukrainian Orthodox Church (UOC), still in communion with Moscow. This raid and seizure were done in the broader national context of "*a "counterintelligence" operation aimed at countering pro-Russian propaganda*", after parishioners were seen at the site singing pro-Russian songs in mid-November. Following the raid in which incriminating materials were found, the UOC's lease on the UNESCO World Heritage Site was terminated, and legislation was passed which would require the UOC to legally identify its name with "Russian Orthodox Church" or "Moscow Patriarchate", which UOC leaders resisted (Palikot 2022).

In responding to UOC complaints, Ukrainian National Security Council secretary Oleksiy Danilov made an appearance on television excoriating UOC leaders for camouflaging ongoing connections to Moscow. Danilov said: "*After February 24, our society needs justice. And if you hide your affiliation with the Moscow Patriarchate, your connections with people who are praying on the territory of Russia for Russian soldiers to kill us, then how can you even look each other in the eye?*" … "*Why are you hiding your name? This is not fair; people should know where they are going. If you are not related, then officially renounce your baptism, say that Putin is Satan and Satan is Kirill. Call a spade a spade, and then it will be fair.*" (Kostyushenko 2022)

Here, Ukrainian security leaders' reciprocal attempts to "desatanize" Ukraine involve the removal of Russian tools of geopolitical influence and an equation of the Russian position to one of evil and not justice.

Following the lead of Ukrainian religious leaders who have portrayed Putin in analogy to both Hitler and the Antichrist, the author argues that Western secular leaders should go on the "spiritual offensive" and be encouraged to make public statements which truthfully articulate the ways in which the behaviour of Russia's leadership during its war of aggression have been "Antichristian" in accordance with Western moral values.

As Alexander Dugin has observed, Putin's rise to power in the "seventh month" of 1999 to emerge as some kind of "ideal ruler of the period" (Dugin 2014) coincides with a popular millennialist Nostradamus conspiracy theory generally interpreted to be about a "Third Antichrist" who would succeed Napoleon and Hitler. (The reciprocal consideration of an "ideal ruler" according to a Ukrainian priest suggests the worldly power of the Antichrist itself (McKay 2022).)

In his respective New Year 1999 and 2000 addresses to Arctogaia, Dugin first reflected on his gleeful anticipation of such a "King of Terror" and then seemed to ruminate that that figure was Putin, and the end times would have to be put off a while (Dugin 1999, Dugin 2000).

There is no doubt that Jesus Christ was not equated with murder or worldly power, and is known as "Prince of Peace", not "King of Terror".

Following this basic "conspiratorial" approach, is easy to conceive of a strategic narrative utilising basic Christian knowledge about Satan, as well as popular eschatological symbols which could depict Vladimir Putin as a usurper of the Orthodox Church and false claimant to the leadership of all Christianity.

Unlike Christ, an ascetic figure associated with life, forgiveness, and truth – who rejected worldly power when tempted by the Devil, the murderous, vindictive, and deceiving figure of "Putin the Third Antichrist" is possibly the richest man on earth, and can be analogised to figures like Pharaoh, Herod, or Nero.

He slays his brothers, poisons his enemies, and threatens the world's destruction with nuclear weapons. Putin can be portrayed as an evil tyrant ("King of Terror") who operates a vast global disinformation machine which "deceives the whole world" (Revelation 12:9).

Such a "strategic conspiracy narrative" based on popular eschatological symbols which contrasts Russian "satanodicy" with Ukrainian (and Western) theodicy (see Epstein (2022)) could theoretically reinforce Western values, while refuting Russia's narrative that its war is Christian.

5. Conclusion

This article applied a semiotic approach to analysis of public statements related to the Antichrist ideomyth in both Ukraine and Russia during the Special Military Operation in 2022. It considered contrasting narratives in the context of a conflict in meanings emerging between the religiously and geopolitically divergent semiospheres of Ukraine and Russia.

The author asserts that this ideomyth has implications for information and cognitive security both in the West and in Russia when framed from the perspective of semiotics and information geopolitics. To this effect, this article has demonstrated that the Antichrist signifier can also be wielded in geopolitics for either offensive or defensive purposes.

Despite the idea that eschatological narratives of Antichrist are Russian symbols which promote cultural unity in opposition to Western values, history demonstrates that another of the significant roles of the Antichrist in Russian semiotics involves political narratives depicting such leaders as Ivan the Terrible, Peter the Great, and Joseph Stalin as Antichrists.

Vladimir Putin would seem to fit in this line, and while in the Ukrainian semiosphere he may have already earned the label of "Antichrist of our current time", it remains to be seen if his symbolic identification with evil would be as globally enduring as that of Hitler.

References

Almond, A. (2020) *The Antichrist: A New Biography*, Cambridge University Press, Cambridge UK

Bagdasaryan, V. (2016) "Идеология как инструмент внешней политики ", *Problem analysis and state-administrative design*, Issue No.2 (46), volume 9, p.6-21

Bagdasaryan, V. & Resnyansky, S. (2016) “Эсхатологические представления в России в дискурсе постсоветских трансформаций”, *Bulletin of the Peoples' Friendship University of Russia*, volume 15, no. 3, p.115-139
Darczewska, J. (2014) *The Anatomy of Russian Information Warfare: The Crimean Operation*, A Case Study, Centre for Eastern Studies, Warsaw Poland
Denysenko, F. (2014) "New Cain" [Online], Ukrainian Orthodox Church Kyiv Patriarchate, https://web.archive.org/web/20140907052044/https://www.cerkva.info/en/messages/5417-new-kain.html
Dolińska-Rydzek, M. (2021) *The Antichrist in post-Soviet Russia: Transformations of an ideomyth*, ibidem Press
Dugin, A. (1999) "С Новым 1999 Годом!", [Online], Arctogaia, https://web.archive.org/web/20200923061654/http://my.arcto.ru/public/Arctogaia-statements/newyear.htm
Dugin, A. (2000) "Новогоднее обращение Арктогеи", [Online], Arctogaia, https://web.archive.org/web/20160608155744/http://arcto.ru/article/1019
Dugin, A. (2014) *Putin vs Putin: Vladimir Putin Viewed from the Right*, London, UK: Arktos.
Dumenko, E. (2022) "Звернення Митрополита Епіфанія (27 лютого 2022 р.)", [Online], Orthodox Church of Ukraine, https://www.pomisna.info/uk/vsi-novyny/zvernennya-mytropolyta-epifaniya-27-lyutogo-2022-r/
Engström, M. (2014) "Contemporary Russian Messianism and New Russian Foreign Policy", *Contemporary Security Policy*, 35:3, p.356-379
Epstein, M. (2022) "Stalin and Satan", [Online], The Times Literary Supplement, https://www.the-tls.co.uk/articles/vladimir-sharov-russian-thinking-essay-mikhail-epstein/
Faure, J (2022) "The deep ideological roots of Russia’s war", [Online], Monde Diplomatique, https://mondediplo.com/2022/04/03ideology
Gallaher, B. & Kalaitzidis, P (2022) "A Declaration on the “Russian World” (Ruskii Mir) Teaching", [Online], Public Orthodoxy, https://publicorthodoxy.org/2022/03/13/a-declaration-on-the-russian-world-russkii-mir-teaching/
Google Trends, https://trends.google.com/trends/explore?date=2022-01-01%202022-12-31&q=%2Fm%2F0g0xk
Halpern, S. (2023) "Putin fires official who called Chabad a 'neo-pagan cult'", [Online], Jerusalem Post, https://www.jpost.com/international/article-729243
Introvigne, M. (2022) "Putin Officially Embraces Theories Accusing the West of “Satanism”", [Online], Bitter Winter, https://bitterwinter.org/putin-theories-accusing-the-west-of-satanism/
Introvigne, M. (2022) "Top Russian Security Official Blames “Cults” for the War, Calls for “Desatanization” of Ukraine", [Online], Bitter Winter, https://bitterwinter.org/russian-official-calls-for-desatanization-ukraine/
Isupov, K. (1995) *Антихрист: антология*, Vysshaya shkola, Moscow
Jansezian, N. (2022) "Ukrainian religious leaders liken Putin to anti-Christ, Hitler", [Online], Jerusalem Post, https://www.jpost.com/christianworld/article-699171
Kadyrov, R. (2022) Telegram (Social Media), [Online], https://t.me/RKadyrov_95/3033
Kostyushenko, O. (2022) "Данилов запропонував УПЦ назвати Патріарха Кирила та Путіна сатаною", [Online], Union of Orthodox Journalists, https://spzh.news/ua/news/70368-danilov-zaproponuvav-upts-nazvati-patriarkha-kirila-ta-putina-satanoju
Krasno, A. (2022), "Пропаганда войны на вере. Помощник Кадырова заявил о "единстве мусульман и христиан"", [Online], Caucasus Realities, https://www.kavkazr.com/a/propaganda-voyny-na-vere-pomoschnik-kadyrova-zayavil-o-edinstve-musuljman-i-hristian-/31899524.html
Laruelle, M. (2022) "What is the Ideology of a Mobilized Russia?", [Online], Russia Post, https://russiapost.info/politics/war_ideology
Lecaque, T. (2022) "Putin’s Violent Holy War Rhetoric Made it to the Christian Right Fringe – And There’s Reason to Believe It’ll Go Mainstream", [Online], Religion Dispatches, https://religiondispatches.org/putins-violent-holy-war-rhetoric-made-it-to-the-christian-right-fringe-and-theres-reason-to-believe-itll-go-mainstream/
Lincoln, B. (1999) *Theorizing Myth: Narrative, Ideology, and Scholarship*, University of Chicago, Chicago
Lotman, Y. (2005) On the Semiosphere. *Sign Systems Studies*, 33 (1):205-229
Madisson, M. & Ventsel, A. (2020) *Strategic Conspiracy Narratives: A Semiotic Approach*, Routledge, London
Magnenou, F. (2022) “Guerre en Ukraine: la "désatanisation" de l'ennemi, nouvelle lubie mystique du Kremlin aux allures de guerre sainte”, [Online], France Televisions, https://www.francetvinfo.fr/monde/europe/manifestations-en-ukraine/guerre-en-ukraine-la-desatanisation-de-l-ennemi-nouvelle-lubie-mystique-du-kremlin-aux-allures-de-guerre-sainte_5445121.html
Maliukevičius, N. (2007) “Geopolitics and information warfare: Russia's approach”, *Lithuanian Annual Strategic Review 2006,* pp. 121-146
McKay, H. (2022) "Ukrainian Church Leaders Join Fight against Putin’s Invasion: ‘He Is the Antichrist’", [Online], National Review, https://www.nationalreview.com/news/ukrainian-church-leaders-join-fight-against-putins-invasion-he-is-the-antichrist/
Medvedev, S. (2022) "Александра Архипова: "Конспирология – это уже новая идеология"", [Online], Echoes of the Caucasus, https://www.ekhokavkaza.com/amp/aleksandra-arhipova-konspirologiya-eto-uzhe-novaya-ideologiya-/32142866.html
MS.Detector.Media (2022) "Хрестовий похід. Як російська пропаганда вигадала «десатанізацію» України", [Online], Media Sapiens, https://ms.detector.media/print/30631/

Orlova, E. (2022) "Спецоперация по кадыризации России, или три союзника Кремля", [Online], Delfi, https://www.delfi.lt/ru/opinions/comments/specoperaciya-po-kadyrizacii-rossii-ili-tri-soyuznika-kremlya.d?id=91730911

Palikott, A. (2022) "Ukraine's Orthodox Church Conflict: A Wartime Struggle For 'Spiritual Independence' And Security", [Online], Radio Free Europe Radio Liberty, https://www.rferl.org/a/ukraine-orthodox-church-moscow-independence/32197737.html

Pertsev, A. (2022) "Александра Дугина много раз называли «мозгом» Кремля. Как утверждают источники «Медузы», его влияние на Путина действительно выросло — но произошло это после убийства его дочери Дарьи Дугиной", [Online], Meduza, https://meduza.io/feature/2022/11/03/aleksandra-dugina-mnogo-raz-nazyvali-mozgom-kremlya-kak-utverzhdayut-istochniki-meduzy-ego-vliyanie-na-putina-deystvitelno-vyroslo-no-proizoshlo-eto-posle-ubiystva-ego-docheri-dari-duginoy

Topor, L. & Tabachnik, A. (2021) "Russian Cyber Information Warfare: International Distribution and Domestic Control", *Journal of Advanced Military Studies,* vol. 12, no. 1

Processing Model and Classification of Cybercognitive Attacks: Based on Cognitive Psychology

Ki-Beom Kim[1,2], Eugene Lim[2] and Hun-Yeong Kwon[2]
[1] Ministry of Defense, Seoul, Republic of Korea
[2]Korea University School of Cybersecurity, Seoul, Republic of Korea

nobody85@korea.ac.kr
eugenelim@korea.ac.kr
khy0@korea.ac.kr

Abstract: Cybercognitive attacks, as witnessed in large and small wars and events along with the recent Russia-Ukraine war, are no longer traditional cyber operations, but are increasingly attacking the psychological weaknesses of targeted members of society and target organizations. Therefore, it is timely to systematically analyse and model cybercognitive attacks. Various definitions and case analyses of cybercognitive attacks are currently being actively conducted, but studies on clear classification and processing models of cybercognitive attacks are almost absent. Accordingly, this paper analyzed cases of cybercognitive attacks. The types derived through case analysis were divided into four categories, and cybercognitive attacks were classified and defined. On such basis, a processing model for cybercognitive attacks was designed, and furthermore, cybercognitive attack layers were classified and presented from the attacker and defender's perspective. The corresponding model and layer presented in this paper model both the countermeasures that can be used to perform cyber operations and the psychological mechanisms hidden in each response process. Specifically, a psychology-based cybercognitive attack processing model was designed to achieve goals by inducing behaviour from collecting information for system managers to inducing response/cognitive processing/decision making/compensation. As such, this paper focused on clarifying the definition of cybercognitive attacks and establishing performance procedures, which are only used as actions using deception by presenting cybercognitive attacks scientifically and logically using psychology descriptions. With that, this paper is expected to serve as the ground for cybercognitive kill chain research that can defend against further cyberattacks using cognitive vulnerabilities.

Keywords: Cybercognitive attacks, human vulnerabilities, Psychology, cybercognitive processing model, cybercognitive hierarchical classification

1. Introduction

The 21st century is witnessing a rapidly changing cyber environment such as the era of the 4th Industrial Revolution and the development of IoT. Meanwhile, cyberattacks targeting national critical infrastructure are frequently occurring, such as Russian-Ukraine cyberattacks, US colonial pipeline hacking, and National Nuclear Security Administration hacking. Despite the fact that National Critical Infrastructure is operated separately from the Internet, the reason for hacking accidents can be cited as 'attacks using cognitive vulnerabilities of system managers', not system vulnerabilities. Thus, research on cybercognitive attacks that exploit the cognitive vulnerabilities of system administrators is required along with research on security technologies of National Critical Infrastructure.

Attacking that exploits human cognitive vulnerabilities in cyberspace is called 'cybercognitive attack', which is not new to modern warfare. British strategist B. H. Liddell Hart randomly selected and analysed 280 wars from the ancient Persian War to the First Arab–Israeli War, stating that only six wars were successful without deception, and that almost all wars were successful using deception. Deception in cyberspace began with the discussion of decoy systems such as honeypots to track intruders in the late 1980s, and cognitive attacks are also described in Tom Kellerman's thesis "Cybercrime Recognition"(Tom, 2017).

Table 1: Liddell Hart Strategy Analysis(19th century – World War II)

Random War Selection	Victory Without Deception Strategy	Victory in Deception Strategy
238 times(100%)	6 times(2.1%)	274 times(97.9%)

Cognitive attack originated from conventional warfare in the past and is applied to modern cyberattack. In particular, cybercognitive attacks can exploit cognitive vulnerabilities of National Critical Infrastructure administrators to access internal networks that are physically impervious.

Accordingly, this paper presents a cybercognitive attack processing model using cognitive psychology for cybercognitive attacks, and classifies the layers of cybercognitive attacks from the standpoint of attackers and defenders.

2. Theoretical Background

2.1 Implications of Cognitive Psychology

Psychology is the study of human and animal behavior and the physiological, psychological, and social processes involved in it. In other words, not only individual psychological processes but also physiological processes that control physical functions, inter-individual relationships and social processes are the subjects of study. Psychology is as old as human history, and Democritus was the first to raise the discussion about whether free will or choice exists, given that our actions are affected by external stimuli more than 400 years ago. As such, research on individual behavior has been conducted through psychology for a long time.

Psychology is closely related to war. During World War I and II, it was implemented in conjunction with deception as a field of battle, and during the Persian Gulf War (1990), it was used as a means of persuasion and public opinion on the international stage, and during the Kosovo War (1998), psychological activities were carried out as a means of cyber. Today, hacking mail, whaling, and honeypot cyberattacks are carried out using the psychology of the target. Psychological warfare is the only weapon that can manipulate the thoughts and actions of the enemy without firing a single bullet. As the continuous nuclear threat grows, cognitive attacks that can manipulate the perception and behavior of targets by means of cyber are the best weapons.

2.2 Proof of Human Vulnerability

Humans are known as rational beings, but based on prospect theory and behavioural economics, humans can be proved as unstable and unreasonable beings.

Prospect theory is an experimental psychology theory that assumes that people are more sensitive to loss than to profit, and that both profit and loss are less likely to feel when they continue to occur, and behavioural economics is the theory of how to behave and what happens by studying actual human behaviour, not rational and ideal economics.

In the light of prospective theory value graphs and behavioural economics, when profits and losses are equal, humans feel more painful when losses occur than when profits occur, and when losses continue to occur, adapting to pain indicates that humans are unreasonable.

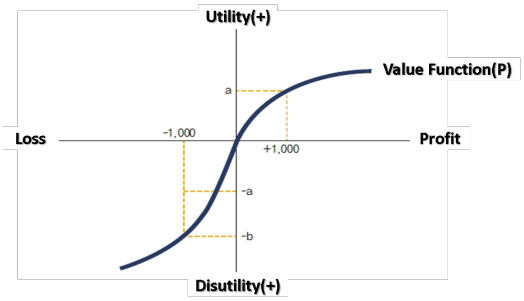


Figure 1: Prospect Theory Value Graph

Prospect theory states that people place more weight on loss than on gain, which means that they place more weight on loss avoidance than on profit-seeking. In other words, it can be inferred that people are much more likely to act not to lose something rather than to act to get something.

If a crisis escalates to avoid losses among countries, the possibility of a preemptive strike may increase. If the current situation is expected to worsen and the loss is expected to expand, policymakers can choose to take preemptive strikes to prevent the expansion of losses(Stein, 1992), and the consequences of avoiding war will further worsen their current status and position(Jervis, 1992).

Looking at the 2003 U.S. attack on Iraq based on prospect theory and behavioral economics, the U.S. reference point is national security and public safety. This is confirmed by U.S. President George W. Bush(2002) claim that war in Iraq is the surest path to peace and security, and Powell's (2002) announcement that collusion between terrorists and WMDs will be the top U.S. concern for years to come. The Bush administration's National Security

Strategy Report, released in September 2002, shows the U.S. government trying to avoid losses through the words ‘protection’ and ‘maintenance’ in its three major goals.

Recognizing the possibility that Iraq could use biochemical and WMD(Weapons of Mass Destruction) against the United States, U.S. President Bush said that Iraq should have a basis for its possible use of WMDs and that if it fails to cope with the threat, something unacceptable will happen (Bush, 2002~2003).

Overall, some view it as an incentive to gain oil-related gains in the Arab region due to the 2003 U.S. attack on Iraq, but the prospect theory explains that it is to avoid loss of U.S. security instability and potential mass destruction of the U.S. people based on national security and public safety.

As such, humans have irrational and cognitive vulnerabilities, and these vulnerabilities are used for cyberattacks.

3. Cybercognitive Attack

3.1 Definition of Cybercognitive Attack

Rosana(2020) classified human cognition into four main categories: Perception, Working Memory, Decision Making, and Action.

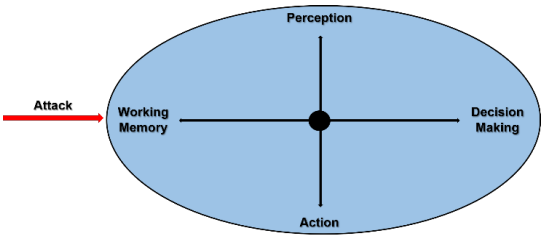


Figure 2: Classification of cybercognitive attack

Perception refers to the conversion of information from words sampled from the senses into neural codes that can be used for intelligent behavior and conscious experience. In the case of Working Memory, it consists of attention and short-term memory, often adjusting processing information by prioritizing specific information over a short period of time to achieve a goal. Decision Making is a gateway to action, prioritizing working memory and information from other unconscious sources. Action means implementing calculations from decision making and other influences. Based on this, attacks on each cognitive element(Perception, Working Memory, Decision Making, and Action) of humans were classified as types of cybercognitive attacks. In this paper, cybercognitive attacks are based on the Working Memory type and are defined as cyberattacks that fuse cognitive psychology to manipulate the perception of targets at the will of the attacker.

Cognitive warfare has been used in the United States since 2017(Underwood, 2017) to describe the ways in which influential groups can act to 'manipulate, infiltrate, influence, or even destroy the cognitive mechanisms of enemies or citizens'. General Goldfein, including James Giordano, a neuroethicist who described the battlefields of the 21st century in the brain and studied the weaponization of neuroscience, described the cognitive war as a war of attrition, while Lieutenant General Stewart of the Defense Intelligence Agency saw it as a cognitive war. The book on Cognitive Superiority, written by Dean S. Hartley et al(2001), developed a theoretical foundation for the sixth area of war that connects technium to noosphere, considered a global expression of human intelligence mediated through technology.

Cognitive attack is a process of changing and deceiving an individual's thoughts and can also be expressed as deception. Deception is defined in Stanford Encyclopedia of Philosophy as intentionally allowing the other party to gain or continue to have false beliefs so that they do not obtain true beliefs. Military flags are defined in the U.S. Military Interpreter JP 3-13.4 MILDEC as "specific actions that allow opponents to deliberately mislead military capabilities, intentions, and operations and contribute to our military operations." Sang et al(2022) defines using defender-dominant information asymmetry as a type of non-cooperative decision-contamination technique that manipulates and deceives potential attackers' cognitive perspectives on defensive deception to continuously construct and maintain false post-action strategies. Defensive cyber deception is viewed as a new concept with unique characteristics that distinguish it from other security factors such as induction, isolation, backtracking, and mutation, while separately having a dedicated Kill Chain process for each application environment and scenario.

3.2 Cybercognitive Attacks Real-world Case

In December 1994, Russia invaded Chechnya to prevent the independence of the Chechen Republic, and the Chechen army fought persistently, conducting cyber propaganda activities such as "about 2,000 soldiers escaped safely despite the siege of the Russian army." Accordingly, Russia blocked cyber propaganda activities by paralyzing the Internet through DDoS attacks. In February 2014, during the Crimean Peninsula War, on the day of the referendum on the annexation of Crimean Peninsula, a website related to the referendum in Crimean Republic was attacked by DDoS, causing it to go dead for about an hour. While Russian cyberattacks were threatening and persistent, Ukraine's response was not clearly visible, so speculation was rampant that it would lose the war with Russia, and that biased broadcasting lowered the rational judgment of the Ukrainian people. In February 22, the Russian-Ukraine war paralyzed government agencies and portal websites, including Ukraine's Defense Ministry, aimed at reducing war fraud, creating fear, neutralizing leadership and resistance, and blocking external information inflows. It can be seen that Russia preoccupied the leadership of the battlefield through cybercognitive attacks and doubled the effect through hybrid operations linked to military operations.

3.3 Literature Review on Cybercognitive Attacks

NATO argued that cybercognitive attacks weaponize public opinion from the outside, causing the target to destroy itself from the inside (Bernal, 2020). Cognitive attacks create destabilization on the public and pursue two complementary goals: government policy and influence on the public. NATO presented six examples of cognitive attacks as follows: confusion-induced destabilization, fragmentation-enhancing destabilization, destabilization as a means of influence, influence recruitment, influence policymaking, and influence as a means of destabilization.

Lucas Hauser(2022) presented a disinformation threat model in which attackers create false information on social media platforms and distribute it to online communities, causing social damage and division. (Step 1: Long-term preparation) False information forms political public opinion, causing division and distrust for future attackers to exploit. In order to spread inflammatory stories, foreign false information providers ask reputable news through followers on social media and try to gain trust in partisan online communities. This behavior(preparation) is the process of gathering information to form the country's domestic political discourse and develop customized disinformation campaigns during the period through cyberattacks. (Step 2: Cyberattack) After deepening the domestic division of the target country, cyberattacks begin at the right time, such as before elections or political scandals. Cyberattacks create a chaotic environment in which customized disinformation can be exploited by damaging computer networks, physical infrastructure, and devices. (Step 3: Tailored Disinformation) Cyberattacks are carried out to escalate the confusion of customized false information. During and immediately after cyberattacks, disinformation channels encourage inflammatory disinformation, maximizing social confusion and polarization and allowing the truth to be buried.

Figure 3: A coordinated Cyber-Disinformation Attack

Johns Hopkins University & Imperial College London classified cognitive operations into single and long-term campaigns. A single campaign focuses on the limited goals of preventing military operations from proceeding as planned or forcing changes in certain public policies, and a long-term campaign confuses entire societies or alliances by pervasive doubts about governance, subverting democratic procedures, and inciting civil unrest or separatist movements. Today, cyber, information psychology, and social engineering are integrated and used to carry out cognitive warfare, and the concept of 'combined weapons' was suggested. It is believed that 'combined weapons' induce behavior that can cause society and groups to become radicalized and divide a cohesive society by sowing seeds of doubt and introducing conflicting negatives.

B.Claverie(2022) evaluate that cognitive warfare is operating around the world as a new war zone following the land, sea, air, space, and cyber areas. This paper describes the concept of cognitive combat, where(aggressive cognitive combat) is a bullying-focused approach and ‘defensive cognitive combat’ involves developing resilience and preventive abilities using similar tools. It also divides cognitive warfare into global perspectives and perspectives based on available tools. ‘Global perspective’ is intended to contribute to a culture of mind coordination or building resilience and global security on the other side of the spectrum and to inform and

educate malicious behavior or intentions, (based on available tools), and considerable interaction with multiple cultures, including decision errors and biases, perceptions and fantasies, cybernetics and control.

3.4 Existing Models and Limitations

Veksler et al (2020) is a model that can generate predictions about the behavior of attackers and defenders, and it is concluded that the threat of cognitive attacks can be reduced by 25% based on small samples of expert decisions using symbolic deep learning (SDL) on the defender's side. He also concluded that the attacker side could use model tracking via dynamic parameter fitting to automatically construct the model during real-time attack scenarios and predict individual attacker preferences by 40-70%. However, the availability of cyber expert decision data is quite rare, often consisting of dozens or hundreds of examples, and DL requires learning data with thousands or millions of examples, so it can be seen that there are limitations. Tzu-Chieh Hung & Tzu-Wei Hung (2020) argued that cognitive attacks and cyberattacks are similar in terms of spreading false information, but cyberattacks differ from cognitive attacks because they attack enemy infrastructure or steal information in practical ways such as distributed denial of service (DDOS) attacks. They also include influencing factors on the human cognitive domain in many types of wars, but there is a limitation in that they concluded that only cognitive warfare weaponizes neuroscience and is committed to brain control. Dit Avocat (2021) provides insightful analysis of how cognitive attacks are implemented on the attacker's side, but there is a limitation that it does not provide details on how human cognition interacts with false information (defenders).

4. Cybercognitive Attack Processing Model

The social engineering cyber operation analysis model presented by Shin et al (2018) modeled the social engineering information collection stage, the social engineering strategy selection stage, and the social engineering strategy execution stage. In the social engineering information collection stage, physical contact with the organization and individual elements to be collected is collected, and in the social engineering strategy selection stage, physical, social, technical, and mixed strategies that can be used socially can be selected. In the social engineering strategy execution stage, it is composed of response of interest, cognitive judgment, and target behavioral stage and psychological agents that operate at each stage.

The social engineering cyber operation analysis model is a model limited to social engineering, and the cybercognitive attack processing model to be presented in this paper is a model designed by collecting and analyzing the following open information that occurred between 07 and 17 years: 20 hacking related to North Korea (Seoul Air Show (ADEX) Defense Contractor Hacking Mail (2015), Blue House Impersonating Hacking Mail (2016), Interpark Personal Information Hacking (2016), etc.), 10 Korean Cases (Hack Mail for Job Applications Targeting Human Resources Officers (2017), Ransomware Impersonating the Payment of Traffic Fines(2017), etc.), 10 overseas cases(The Dalai Lama's Computer Hacking (2009), London Olympics Official Site Impersonation Scam (2012), WhatsApp Fraud (2016), etc.).

Once the information on the attack target has been collected, a response is induced with the following four factors: Expertise which corresponds to the target's specialized field and is recognized as high-quality information, Salience that gives a noticeable stimulus such as color, size, sharpness and movement of information, Curiosity to be interested in celebrities, financial gain and sex-related contents, and Urgency such as information with an ‘urgent’ prefix or with limited access and processing deadlines. It is also expressed as a social engineering technique. If it succeeds here, it will move to the victim's territory, and if it fails, it will be re-performed from collecting information.

Cybercognitive attacks provide false information to the target's cognition, and in this process, cognitive processing and schema are formed to compare and analyze knowledge and behavioral types of the external world stored in the individual's head. Cognition maintains a state of equilibrium, and there is a assimilation process in which false information is accepted and generalized in line with existing schematics, or a control process in which existing schematics are changed or newly made occurs when they do not fit the existing schematics. Finally, there is an instinctive process of organization and equilibrium that reorganizes existing schematics into new and complex structures. After schematic processing, decision-making is made using psychological descriptions related to decision-making such as speculation, partial thinking, group thinking, internal group bias, and prospect theory. If the victim is rewarded with the information or results he or she wants for subsequent actions, he or she will behave as the attacker manipulated without reasonable doubt.

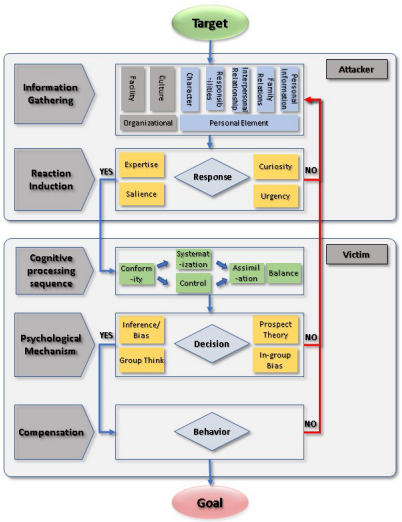


Figure 4: Cybercognitive Attack Processing Model

5. Cybercognitive Attack Hierarchical Classification

Until now, studies related to cybercognitive attack have been focused on cognitive aggressive means, procedures, and cases, and classification according to the type of technology that occurs in the process of human cognitive attack has been sluggish. Accordingly, based on the cybercognitive attack processing model, hierarchical classification was performed on cybercognitive attacks that can be performed from the attacker's point of view and the defender's point of view.

5.1 Cybercognitive Attack from an Aggressive Perspective

Attacks can neutralize the cyber defense system by exploiting the cognitive vulnerabilities of the system personnel responsible for the internal network. An attacker can intentionally provide false information to the victim to induce victim behavior and provide a contact point for the network to penetrate the disconnected closed network.

5.1.1 Spear Phishing

Spear phishing is a highly targeted attack that sends hacking emails with sophisticated design based on the names of people involved, words in company documents, information that makes you believe in normal mail, and common items at work. FireEye analysed that 91% of cyberattacks started from hacking mail, 99.7% of attachments used in hacking mail use social engineering techniques, and 65% of spear phishing attacks were successful. Symantec's 2019 Internet Security Threat Report(ISTR) shows that the number one word in the email attack's title is "urgent". As such, spear phishing attacks are attacks that combine psychological mechanisms based on the information of the target.

5.1.2 Fake News

Fake news causes individuals to believe inaccurate information and influence their opinions and actions. In the past, various groups of people, such as advertisers, political activists, and religious fanatics, used various forms of fake news to influence public opinion or spread propaganda. And modifying fake news doesn't necessarily change people's beliefs. Even if you believe in the revised truth, fake news continues, and the act of refuting it can have negative consequences. The more people hear fake news, the more familiar it becomes, and the more likely it is to believe it real. This effect is due to familiarity and fluency bias.

5.1.3 Reverse Social Engineering

Social engineering is a psychological technique that tricks a person into moving in the direction they want, i.e., by deceiving them to leak sensitive information or bypass the security perimeters associated with information, Reverse Social Engineering(RSE) is a technique that induces an attacker to perform a malicious attack, such as a direct attack. In the past, espionage agents established friendship by solving the problem that the target was in, formed a rapport, recruited, and ordered espionage.

5.2 Defensive Cyber Deception Operation

Rather than waiting for an enemy in defense, you can actively bait them to reduce risk by exploring and attracting them. Defenders provide the attacker with false information that they accept as true, disrupting the attacker's asymmetric advantage and making them believe incorrectly It also has the advantage of being able to get a

counterattack opportunity and reduce the possibility of attack and cyber defense costs by identifying the attacker's strategy, tactics, capabilities, and intentions.

5.2.1 Honey Pot

Talinn Manual 2.0 defines Honeypot as a trick to protect computer systems from malicious manipulation using a physical or virtual environment designed to attract intruders' attention with the aim of deceiving them. Honeypot creates an independent, vulnerable environment that induces malicious users to connect, making it seem like it can be used to steal information from that system or attack other systems. Most of the actions inside the Honeypot are limited, and the commands you enter remain in the log, and the downloaded files are stored separately on the host system. It then records and analyzes all actions of malicious users to respond to subsequent attacks. Security policy design and content are important for honeypot operations. A security policy that is too low can cause an attacker to suspect or avoid access. And the content of the framework needs to be updated continuously.

5.2.2 Cyberattack Traceback

Traceback is a technology that can track the source of attacks. Traceback technology includes Host-based TCP Connection Traceback, which uses backtracking modules installed on systems to locate attackers via multiple different systems, and IP Packet-based Traceback, which installs backtracking modules where network packets can be monitored. Analysis of collected data using Traceback technology can identify attack environment information such as attacker's IP, country, operating system, monitor size, and find attack patterns such as activity time, collected information, and attack tools. From the defender's point of view, you can analyze the attacker's attack environment and pattern to get a chance to counterattack.

5.2.3 Cyber Persona

People represent themselves as cyber personas in cyberspace. Cyber personas form a layer with the physical and logical layers of cyberspace. Cyber personas allow attackers to gather information or identify system weaknesses, but can also be used in reverse.

In the cyber space where the attacker is active, cyber persona can be used to find out the attacker's ability, personality, and interests. IT personnel from North Korea's Kim Il-sung University also team up with athletes from other countries at the CodeCodeforce. If you have formed a close relationship through team activities, you can leak false information such as honey doc to induce attacks or lower rational judgment. Recently, Metaverse, a three-dimensional virtual person working in cyberspace, has not yet been clearly established, but it is being used as a world where real life, legally recognized activities, jobs, and financial learning are connected in three dimensions. Metaverse is an opportunity and a challenge for the IT industry, and there is a risk of hacking such as biometric information hacking, brain hacking, random thought injection, and behavioral pattern manipulation.

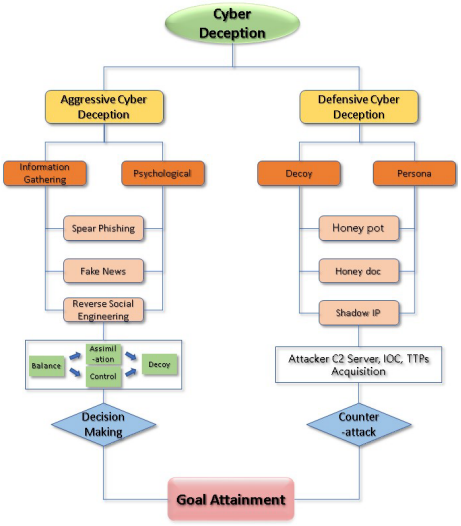


Figure 5: Cybercognitive Attack Layer Classification

6. Conclusion

The cause of malicious code infection targeting the national important system has been determined by receiving hacking mail and social engineering techniques, and technical analysis such as TTPs, source codes, and attack patterns of malicious codes has been conducted. As a result, malware response technology and security systems have become more advanced and robust. Nevertheless, hacking accidents occur because the attacker is using the cognitive vulnerability of the system manager, not the attack using the system vulnerability.

Identifying and attacking systemic vulnerabilities requires a lot of time, budget, and risk, such as zero-day vulnerability development or purchase, supply chain attacks, and physical penetration. However, system managers' cognitive vulnerabilities can benefit from reduced budget and effort, and system incapacitation that cannot be physically infiltrated. In order to attack important national systems, attacks using system manager cognitive vulnerabilities are effective, and cybercognitive attacks are performed because the probability of attack success can be increased.

Improving the security level of the national critical system requires research on technical responses and cognitive attacks using system managers. Previously, system managers judged and processed the path of malicious code infection with simple social engineering techniques, or classified according to the type of technology that occurs in the process of human cognitive attack through cybercognitive attack methods, procedures, and cases. This paper presents a cybercognitive attack processing model based on psychological response induction, psychological writing, and defender cognitive processing based on psychological psychology that scientifically studies decision-making after dividing into attacker and defender areas. In addition, it is meaningful in that it promoted hierarchical classification of cybercognitive attacks and defense activities in cyberspace, and the definition of cybercognitive attacks was clear and the performance procedure was established.

If additional research is further conducted on the attacker's cognitive processing process and purpose of attack in the cybercognitive attack model, it will be possible to devise strategies such as cyberkill chains that can identify and prevent/reject cybercognitive attack mechanisms.

References

Al Amin, M., Shetty,(2019). Attacker Capability based Dynamic Deception Model for Large-Scale Networks.*ICST Transactions on Security and Safety*,*6*(21). https://doi.org/10.4108/eai.13-7-2018.162808

Allcott, H., Gentzkow, M.(2017). Social media and fake news in the 2016 election. In *Journal of Economic Perspectives*(Vol. 31, Issue 2). https://doi.org/10.1257/jep.31.2.211

Bernard Claverie, Baptiste Prébot, Norbou Buchler, François du Cluzel(2022), What Is Cognition? And How to Make it One of the Ways of the War?, NATO Collaboration Support Office, pp.4, 1-17, 978-92-837-2392-9.

Bernal, A., Carter, C., Singh, I., Cao, K., Madreperla, O. (2020). “Cognitive Warfare: An Attack on Truth and Thought”, NATO and Johns Hopkins University: Baltimore MD, USA.

Chang, Keung Ryong.(2004). Why did the U.S attack Iraq?: An Explanation based on Prospect Theory pp. 89-112. The Journal of Political Science & Communicaiton

Flynn, D. J., Nyhan, B., Reifler, J.(2017). The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics. *Political Psychology*, *38*. https://doi.org/10.1111/pops.12394

GRAPHUS .(2020). Retrieved from https://www.graphus.ai/resources/spear-phishing-social-engineering/

Greenhill, K. M., & Oppenheim, B.(2017). Rumor has it: The adoption of unverified information in conflict zones. *International Studies Quarterly*, *61*(3). https://doi.org/10.1093/isq/sqx015

Hasher, L., Goldstein, D., Toppino, T.(1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1). https://doi.org/10.1016/S0022-5371(77)80012-1

Hung, T. C., & Hung, T. W.(2022). How China’s Cognitive Warfare Works: A Frontline Perspective of Taiwan’s Anti-Disinformation Wars. *Journal of Global Security Studies*, *7*(4), 1–18. https://doi.org/10.1093/JOGSS/OGAC016

Jervis, R. (1992). Political implications of loss aversion. Political Psychology, 13(2), 187–204. https://doi.org/10.2307/3791678

Kim, J.T., Han, M.H., Lee, J.H., Kim, J.H., & Kim, I.K.(2014). *Technical Trends of the Cyberattack Traceback*(pp. 93-103). Electronics and Telecommunications Trends Volume 29 Issue 1: Electronics and Telecommunications Research Institute.

Kulíšek,J.(2012). Military Deception. *Vojenské Rozhledy*, *21*(2). https://doi.org/10.3849/2336-2995.21.2012.02.040-058

Lee, Jonghyun.(2018). Critical Analysis on the Application of Due Diligence Principle in the Cyber Context. *Korea International Law Review*, *49*, pp. 143-170.

Liddlell Hart, B. H.(1967). Strategy second revised edition. In *International Affairs*.

Mary-Ann Russon.(2021). *US fuel pipeline hackers 'didn't mean to create problems'*. BBC News.

Murray, A.(2016). *How to report fake news to social media* , BBC News.

Myers, D. G., & Dewall, C. N.(2015). Psychology in Modules - by Myers and DeWall. In *Worth Publishers*.

Natasha B, Eric W.(2020). *Nuclear weapons agency breached amid massive cyber onslaught*. Politico.

Norman, D.A.(1986), Cognitive Engineering, In D.A. Norman & S.W. Draper, Eds. User Centered Systems Design, pp. 15-34, Hillsdale, NJ: Lawrence Erlbaum Associates

Oh, H.(2015). Unsupervised Scheme for Reverse Social Engineering Detection in Online Social Networks. *KIPS Transactions on Software and Data Engineering*, *4*(3). https://doi.org/10.3745/ktsde.2015.4.3.129

Piaget, J.(1932). The moral judgment of the child. London: Routledge & Kegan Paul.

Sang Seo, Dohoon Kim, Sangwoo Han(2022). Simulation and Analysis of Drone Combat Damage Effect Against Cyber-Electronic Threats. *JKIIT*, 20(11), 163-185.

Schmitt, M. N.(2013). Short form citations. In *Tallinn Manual on the International Law Applicable to Cyber Warfare*, Rule 6, para. 28, Cambridge: Cambridge University Press.

Stein, J. G., & Pauly, L. W. (1991). Choosing to Co-Operate: How States Avoid Loss. International Journal, 47(2), 199–201.

Shin, D. C., & Park, Y. H.(2017). Development of Risk Assessment Indices for Social Engineering Attacks. *Journal of Security Engineering*, *14*(2). https://doi.org/10.14257/jse.2017.04.01

Shin,K., Kang,J., Yoo,J., Kim, J., Kang, S., Lim, H., & Kim, Y.(2018). A Study on the Concept of Social Engineering Based Cyber Operations. *Journal of The Korea Institute of Information Security and Cryptology*, *28*(3).

Stoll, C., & Connolly, J.W.D.(1990). The Cuckoo's Egg: Tracking a Spy Through the Maze of Computer Espionage. *Physics Today*, *43*(8). https://doi.org/10.1063/1.2810663

Wells, L.(2017). Cognitive-Emotional Conflict: Adversary Will and Social Resilience. *Prism*, *7*(2).

Yoon Soo, Choi.(2015). *Results of interim investigation into KHNP's cyber terrorism case*, Republic of Korea: Prosecution Service.

Designing an Email Attack by Analysing the Victim's Profile: An Alternative Anti-Phishing Training Method

Dimitrios Lappas and Panagiotis Karampelas
Hellenic Air Force Academy, Dekeleia, Greece

dlappas.aegean@gmail.com
panagiotis.karampelas@hafa.haf.gr

Abstract: According to Thomson-Reuters the top cyber threat today is phishing in which people are tricked either to click a malicious link or give out personal information. It's a fact that 96% of these phishing attacks comes from emails, which amount to more than 3.4 billion daily, as reported by Cisco. Austrian aerospace company FACC, Belgian bank Crelan, Acorn financial services and many other companies were recently fell victims of phishing emails losing millions of dollars. Even if experts provide lists of signs that users should seek in an email in order to understand if it is legitimate or scam, the attackers have elevated the quality of the email messages making them believable and very hard to discern them. In order to respond to this elevated threat, unconventional user training is required, focusing on recognizing a phishing email. Knowing how an attacker thinks and prepares the attack vector against a target, we claim that it will make users more suspicious when they receive one. In this regard, an innovative education intervention (consisted of two phases) was designed and developed. In the first phase, 98 participants were asked to visit an artificial social media profile and prepare a phishing email in order to persuade the victim to click a link. Then, the participants were presented with an innovative guided workflow to prepare a spear phishing email which was based on social media intelligence. In the second phase, they were asked to prepare one more email for the same person applying this time the guided workflow. Comparing the two different emails created, we found that the guided workflow led to the creation of more authentic emails which could potentially trick the victim easier. Based on the theory of active learning, we believe that by teaching users how attackers exploit their personal information in order to develop their attack vectors, it will increase their awareness not only for the typical phishing emails but also for more sophisticated spear phishing attacks.

Keywords: Phishing, Spear Phishing, Email attack, Anti-phishing training.

1. Introduction

According to Thomson-Reuters[1] in 2022, the top cyber threat today is phishing in which people are tricked either to click a malicious link or give out personal information or provide their credentials to enterprise systems. Based on the statistics of earthweb[2], more than 3.4 billion phishing emails are sent daily which accounts for the 94% of all the cyber-attacks that take place globally in a day and for the 90% of ransomware attacks (Branca, 2023). In addition, it is estimated that more than 1.5 million new phishing web pages are created every month usually imitating some of the most frequently used web pages of social networks, e-banking web pages or web email services with the purpose to lure unsuspicious users to provide their credentials in the respective service. What is very impressive according to the statistics is that 20% of the employees who will receive a phishing email, they will click the link or they will open the attachment (Branca, 2023) and the 67.5% of the individuals who click the phishing link to the fake website, they will continue and provide their credentials[3]. However, email is not the only mean used for phishing attacks. Another similar attack is smishing which is the use of SMS text messages to trick users to provide most of the times their user credentials. Recently, a more interactive attack has appeared called vishing. In this attack, there is a phone call in which the caller attempts to trick the receiver to provide confidential information over the phone and thus get access to bank accounts or other personal assets. Angler phishing is another one that takes place in social media in which persons pretend that they belong to customer services and they are approaching disgruntled customers asking, during their message interchange, personal information. Pop-up phishing appears in legitimate websites which have been infected with malicious code and present pop-up windows asking for the user credentials pretending that this is on behalf of the legitimate website. Another phishing attack is whaling when the target of the attack is a senior executive member of an enterprise and finally spear phishing in which emails are used with the difference that the phishing email sent to the target is well-prepared and most of the times personalized which makes it believable.

The impact of the phishing attacks is estimated to be around 4.5 million dollars per attack especially when data breaches are associated with the successful phishing attack which can be even higher in case of spear phishing

[1] https://legal.thomsonreuters.com/en/insights/articles/top-5-most-common-cyber-security-threats-today
[2] https://earthweb.com/how-many-phishing-emails-are-sent-daily
[3] https://www.cisco.com/c/dam/en_us/about/doing_business/trust-center/docs/phishing- program-infographic.pdf

attacks (Branca, 2023). A prominent example of such spear phishing attacks comes from the Austrian aerospace parts maker, FACC, in which a fake email supposedly coming from the Chief Executive Officer of FACC, Walter Stephan, requested an employee of the accounting office to transfer 47 million dollars to an unknown bank account for a fake acquisition project of the company[4]. A similar scam hit the Belgian bank Crelan in which the attacker imitated the CEO of the bank and in an email asked an employee to transfer approximately 75 million dollars to an unknown account[5]. Recently, in April 2022, the financial services company Acorn uncovered a major data breach in which the attackers had managed to get access to personal information of Acorn's customers such as name, address, date of birth, driver's license number, financial account number, Social Security number, and other account-related information[6]. The attackers manage to compromise the internal network of the company by stealing the email credentials of an employee via a phishing email sent to him (McCurdy, 2022).

According to Gupta et al (2017), there is a plethora of defence techniques against phishing attacks. One prominent category is cultivating the awareness of the user through training while the alternative one is based on automated software detection techniques. These techniques can be applied in the network level in which known IPs addresses that are associated with phishing websites are not allowed in the network. Alternatively, authentication-based mechanisms can be applied in email communication such as digital signing of each email which confirms the origin of the email message. Another technique is to activate client-side tools that are able to detect phishing emails using blacklists and other similar services and finally, server-side tools can be installed that are able to detect phishing emails using machine learning techniques before the emails reach the recipient. Even if some of the automated methods are able to detect the phishing emails with high precision, the attackers are always a step ahead and can easily advance their techniques and thus trick the detection systems as it happened in Acorn case (McCurdy, 2022). It is widely accepted (Gupta et al, 2017; Roepke et al, 2020; Sumner et al, 2021) that the only way to minimize the risk and the loss coming from the phishing emails is to empower users' awareness by applying the appropriate anti-phishing training techniques.

Usually, the anti-phishing training programs are focusing on lists of features that users should seek in an email in order to understand if it is legitimate or scam and this type of training is usually delivered either as a self-paced game (Sheng et al, 2007) or in a structured training session (Dodge et al, 2012). However, the attackers have elevated the quality of the email messages making them more believable and thus it is very hard to discern them. In order to respond to this elevated threat, unconventional user training is required, focusing on recognizing a phishing email not only by its technical features but also by the phishing tactics that are used e.g., emotional manipulation, perception of need, cunning communication or perception of need[7].

We claim that, knowing how an attacker thinks and prepares the attack vector against a target, it will make users more suspicious when they receive a phishing email and so they will not be tricked. In this regard, an innovative anti-phishing education intervention is proposed in order to empower the awareness of the email users. Our approach, is based on the experiential learning theory and consists of two phases, including a hypothetical scenario for practice. Our method aims to train the students on how to prepare a spear phishing email using a structured methodology. We claim that using the proposed educational intervention, students have the opportunity to:

- improve their understanding and perceptions around the protection from phishing emails
- increase their self-confidence in the area of reconnaissance a phishing email, as they will be able to construct one
- apply the theoretical knowledge of phishing email deception
- reflect on the tactics an attacker can use in order to reach his/her goal.

The paper is organized as follows: section 2 presents the related anti-phishing awareness training programs. Section 3 presents in detail our training methodology. The next section focuses on the assessment of the proposed method in the context of an international course and the relevant results are presented. Section 6 discusses the findings and the limitations of the proposed approach and the final section concludes the paper with the relevant results.

[4] https://legal.thomsonreuters.com/en/insights/articles/top-5-most-common-cyber-security-threats-today
[5] https://www.pindrop.com/blog/ceo-phishing-scam-costs-belgian-bank-crelan-75m
[6] https://www.jdsupra.com/legalnews/acorn-financial-services-reports-data-5996771/
[7] https://www.microsoft.com/en-us/security/business/security-101/what-is-phishing

2. Related Work

As phishing has become one of the top cyber threats the recent years, anti-phishing training has received a lot of attention from the security community in order to raise users' awareness either by informative messages or by specialized training regarding the potential counter-measures that should be taken in order to eliminate its ruinous effects and the respective cost (Kumaraguru et al, 2007). Anti-phishing training most of the times appears in the form of (a) simulation-based training by sending mock phishing emails to the trainees with the task to distinguish the genuine emails from the phishing emails and (b) game-based training in which trainees are called either to identify phishing emails and URLs or to construct through an interactive game a malicious link which could be used in a phishing email (Roepke et al, 2022; Vayansky and Kumar, 2018; Abbasi et al, 2016). Although there are several educational game applications which attempt to address the issue, there is still a gap in anti-phishing training as it will be presented.

A first approach, in the context of situated learning, presented a comic strip intervention when emails users received a training email (Kumaraguru et al, 2007). The comic strip showed how the users can recognize a phishing email by examining various elements of the email, i.e., the sender and the embedded link and what they should do when they suspect that the email is a phishing one.

A mobile game has been proposed in which "a small fish wants to eat worms in order to become big fish" but not all the worms are safe since many of them contain illegitimate URL addresses (Arachchilage and Cole, 2011). The specific game focuses on teaching email users to distinguish malformed URLs and thus avoid clicking them. Similarly with the relevant automatic classification techniques, the approach is not so effective since there are phishing emails that instead of URL addresses they use file attachments as bait.

Another mobile application game was designed with the purpose to introduce to the players the features that characterize a phishing email. The game had 10 levels with increasing difficulty, presenting in each level an introductory story and then a relevant exercise. Overall, the game was focusing on how players would be able to recognize a fake URL which is not structured correctly or had only an IP address or used derivative domain names, etc. (Canova et al, 2014). However, other phishing features were not mentioned, such as file attachments or fake senders, etc.

Another anti-phishing game, called PHREE OF PHISH, was designed by Pars (2017) in which the player assumes the role of a web shop employee in which attends four lessons regarding features of phishing emails. Then the player is called to respond to time-constrained multiple questions regarding phishing emails. The first lesson presents generally the tactics of the attackers, the second one presents the characteristics of the phishing emails in which the user should pay attention such as the sender of the email, the links embedded and the content of the message if implies time pressure, asks for personal information, includes threats or awards and contains typos. The third and fourth lessons focus on the characteristics of the URLs that could be embedded in a phishing email. While the game covers a wider range of features of phishing emails, it is mainly focused as the other games on how the user can recognize the fake URL and not to all the potential tactics an attacker could use.

In the What. hack, the authors designed a game in which the player assumed the role of an email controller who was responsible to check an email arrived in a company and decide whether it is legitimate or fraudulent. The game put the players under pressure since in eighty seconds they had to process six emails going through the five levels of the game application. In each level, different phishing features were examined such as trusted/untrusted domains, unknown domains, suspicious URLs and file attachments (Wen et al, 2019). While the game seemed to be effective in anti-phishing training, it did not address the spear phishing features of the respective emails which mainly are found in the origin and content of the email.

In another interesting study, the authors created a system to automatically harvest from a social network the community of friends of students in their University. They, subsequently, sent short phishing emails to all students faking that the sender was one of their friends that they found in the social network. The result of this experiment was that the possibility to click a bait link in a phishing email is four times higher when the email sender impersonates a friend of the victim (Jagatic et al, 2007). The specific outcome of the study in all the previously presented anti-phishing trainings is not taken into consideration even though it is one of the main features of a spear phishing attack. Based on the findings of the specific study and in an effort to make the email users aware of all the potential features an attacker can use in phishing emails, we propose an innovative learning approach which trains the participants how to design an effective spear phishing email enabling them at the same time to reflect on the tactics of an attacker, increase their confidence regarding the detection of

phishing email and apply the theoretical knowledge of phishing email deception as it will be presented in the next sections.

3. Training Methodology

3.1 Experiential Learning

The process of acquiring, retaining and using appropriate information constitutes the phenomenon of learning (Estes, 1975). As a result of the process of learning is the change of human behaviour (Gagné, 1975). Looking through the literature, one will come across with many theoretical approaches of learning since many suggestions have been made on how instructors will be more effective and how learners will achieve the desired learning outcomes. Some examples of learning approaches are the behaviourist approach (Watson, 1930), constructivist theory (Bruner, 1960), sociocultural theories of knowledge (Vygotsky, 1978), exploratory (Bruner, 1960), cooperative (Golub, 1988) and experiential learning (Evans, 1994).

Experiential learning focuses on the development of experience during the learning process, through observation, reflection and action. It utilizes students' experiences (already existing experiences) and elicits new experiences, as opposed to memorizing information (Evans, 1994). In this context, instructors have to create an appropriate environment that will lead trainees to the discovery of knowledge through experiential learning. On the other hand, students should be encouraged to actively participate in the learning process according to their personal interests. One such educational process is the methodology of active learning which engages learners through their participation in certain tasks within a specific theoretical framework. (Meyers & Jones, 1993; Bonwell & Eison, 1991).

Active learning focuses on building experience rather than just remembering the theory and for this reason students are engaged in activities in which they have the opportunity to develop their skills actively (Bonwell & Eison, 1991; Coulshed, 1993; Felder & Brent, 2003). Simulation of problem-solving cases and role playing are two keys approaches in active learning (Cash, 1983). Instructors take the role of facilitators in the learning process by organizing students' activities and help them in case of further explanation. Every mistake is used by instructors as a chance for discussion and explanation in the whole learning team. Thus, learners are encouraged to try anything new or unusual as far as they are in the right learning track.

Scenarios, interactive or non-interactive, are used as basic tools of the teaching and learning process (Clark, 2009). Active learning strategies are also supported by scenarios and students are provided with the opportunity to learn and apply their learning to real-world experiences (Errington, 2005). More specifically, students work through a story based on a problem situation that they are required to solve. In other words, students apply their knowledge, critical thinking and problem-solving skills in a safe environment that resembles the real world (realistic learning environment).

The choice of an appropriate scenario can help the learning objectives to be achieved in two ways (Margetson, 1998): either as the story (model example) in which the acquired knowledge will be organized, or as the story that will help in the practical understanding of the knowledge. The second one consists the common place of scenario-based learning and active learning. According to Thomsen et al (2010) learning takes place experientially with the active participation of students in solving a problem under the guidance of the teacher.

3.2 The design of our learning approach

Our learning approach was based on the experiential learning theory and was designed in two phases, including a hypothetical scenario for practice.

Phase 1

The first phase of our course included a lecture about social engineering and a practical component in which the students had to construct a phishing email. During the lecture, attacking methods and techniques were presented that are used in order to steal passwords or to become trustworthy to the victims in order to extract from the victims confidential information. Then, the students were asked to visit a hypothetical person's profile on social media (Figure 1), by using their smartphones, and create a phishing email for him. There was no guidance on how a phishing email can be constructed but they had to think creatively based on what they had heard during the lecture.

Phase 2

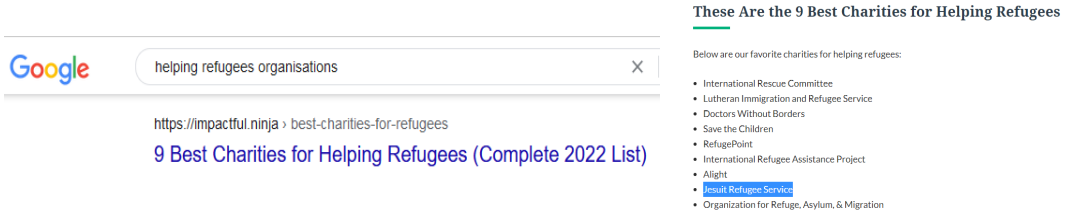
The second phase of the course included a lecture focused on how a social engineering attacker thinks and acts. An innovative method was presented describing the steps that the attackers may follow in order to make victims to believe their message and click on a suspicious link. Several real life examples were also presented in which this methodology would work. Analytically, the proposed methodology included the following steps:

- Step 1: Information gathering about the victim

  Searching on Facebook, we found a team called "Refuges Welcome – Greece" (Fig. 1a). In a post about "Migrants", there was a comment of a user (Fig. 1b). Then, we found this user on Facebook and her friends (Fig. 1c).

(a) (b) (c)

Figure 1: Excerpts from Facebook in which the original post is visible (a), the response of the victim in a comment (b) and the friends of the victim (c)



(a) (b)

Figure 2: The results from the relevant search in Google

- **Step 2:** Grouping – categorization of the victim's information

  Based on the content analysis of the message we extracted the following information:

  - Her grandparents lived a war in Europe
  - Her grandparents survived from the war and rebuilt Europe
  - She believes that the same must be done for Syrian refugees nowadays

- **Step 3:** Evaluation of information - choose to use the category in which the victim expresses more liking or spends more time.

  Based again on the content analysis the author expresses the opinion that:

  - She believes that Syrian refugees must stay in their country

- **Step 4:** Search for dominant emotion or desires of the victim in the above category

  In the content of the message the author expresses a desire:

  - She believes that people who have survived from a war must rebuild their country

- **Step 5:** Mapping the victim's contacts – friends who have the same interests

Examining the contacts list, useful observation can be made, as for example:

- A friend of her has faith to Jesus (Fig. 1c in the center of the first row)

- **Step 6:** Finding a real or imaginary entity (organization or person) that the victim could trust
  - Searching on Google the phrase "helping refugees organizations", we found a list with charities and one of them is the Jesuit Refugee Service (Fig. 2)
- **Step 7:** Finding a motive for the victim to follow the instructions included in the e-mail

  Synthesizing the abovementioned findings, we have the following modelling:
  - Morphological Analysis of the situation
    - ✓ Her grandparents lived a war in Europe
    - ✓ Her grandparents survived the war and rebuilt Europe
    - ✓ She believes that the same must be done for Syrian refugees nowadays
    - ✓ She has a friend with great faith in God
  - What if we made a message including:
    - ✓ Refugees of Syria needing help
    - ✓ It is better to help them rebuild their country than become refugees
    - ✓ Jesuit Refugee Service has already started a similar campaign
    - ✓ Her friend gave them her email
- **Step 8:** Views of experts
  - Searching on Google for expert's opinion about refugees, more specifically about quotes, we found that Amela Koluder claimed that "*a refugee is someone who survived and can create the future".*
  - Searching on Google pictures with the phase "Jesus about refugees", we found an image pointed that "*Jesus was a refugee too (Matthew 2:13-19)"*

Finally, based on the proposed methodology, we constructed an example of a spear phishing email, as it appears in Figure 3 (it is supposed that by clicking the link a malware will be transferred to the device)

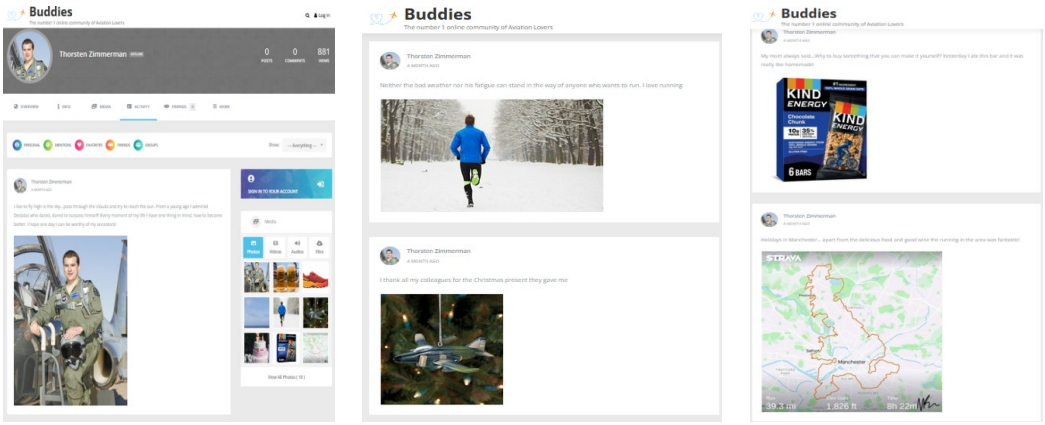
Dear (User Name),
My name is L.O. and I am a volunteer nurse of the Jesuit Refugee Service. I have spent almost all my life in a refugee hospital and I feel sad and angry when I see Syrian people suffering from the war. Unfortunately, many of my patients die or become refugees to Europe. But personally, I believe that if we really want to solve this problem we must help these people stay in Syria and rebuild their beautiful country when the war will stop. As Amela Koluder claimed "a refugee is someone who survived and can create the future".
Your friend (username) suggested me to contact you since you share a similar attitude towards the refugees. If you want to participate to our effort, to help Syrian people stay in their country and not to become refugees, please donate to the Jesus church of Syria. Here is the relative link: http://.....
Don't forget that Jesus was a refugee too (Matthew 2:13-19), so he can feel us better!
Thank you in advance
L.O.

Figure 3: The spear phishing email that was used as an example

After the presentation of our proposed methodology and the corresponding example, trainees were asked to follow these steps and make a new message of a spear phishing email for the same hypothetical scenario used in phase 1.

The hypothetical scenario for practice

The scenario for practise in both phase 1 and 2 included a hypothetical person who had a profile on a social media page. On this page, he posted details about his life including job, hobbies, family, friends and likes. More specifically, the hypothetical person was a German pilot of a fighter aircraft who loved running and he was very keen on the military history of aircrafts. He had special healthy diet preferences and he liked specific running shoes. He was also referring to many posts to a close friend of him named Josh and to his daughter. Everyone could access his profile in the designated social media (Fig. 4).



(a) (b) (c)

Figure 4: The fake profile that was used for the practice

Participants took the role of a social engineer. Their task was to create a phishing email. More specifically, they were asked to create the content of a phishing email and send it to this hypothetical person. The objective of the email was the victim to click on a hyperlink which automatically would download a malware in his device with the purpose of collecting confidential information.

4. Experiments & Results

The course took place in the Hellenic Air force Academy (HAFA) on the 3rd of November, 2022. It was completed in 7 teaching hours as part of one week module. The title of the course was "Syndicate Work – Social Engineering" while the title of the module was "Cyber warfare". This module was part of the analytical curriculum of the International Air Force Semester in the context of a Military Erasmus program. The course took place in the main auditorium of HAFA.

Participants attended the course came from international military academies and civil universities. More specifically 98 students and cadets attended the course. Military cadets came from Air Force Academies of Greece, Portugal, Romania, Italy, Spain, Poland and Austria, while students from civil Universities came from Greece. The age of the participants was spanned from 19 to 23 years old and there were 67 men and 31 women. All the participants had attended three days of lectures related to "Cyber warfare" before attending our lecture.

Participants as mentioned, had to create a phishing email for our hypothetical scenario two times targeting the same individual. In phase 1, they had no guidance on how to make it and in phase 2, they had to follow the steps of our proposed methodology for the construction of the phishing email. To evaluate our proposed methodology, we compared the messages that the trainees created in phase 1 with the corresponding ones in phase 2.

Focusing on the elaboration of the contents of the emails, we compared the average words and sentences of their emails (Table 1). It is observed that during phase 2, the email content was further processed since on average more words and sentences were used than in the first one.

Table 1: The results of the experiment

Metrics	Phase 1		Phase 2	
Average words per email	50,57		65,29	
Average sentences per email	5,52		6,47	
Emails that didn't include any detail of the victim's profile	14 out of 98	14.29%	3 out of 98	3.06%
Trainees used the job of the victim as the main subject of their email	43 out of 98	43.88%	32 out of 98	32.65%
The message of the email was personal and it was using the victim's name	46 out of 98	46.94%	68 out of 98	69.39%
The email included references to friends	5 out 98	(5.10%)	36 out of 98	36.73%

The second factor that was examined was how the social media profile of the hypothetical person helped trainees to form their email. In that sense, we compared the number of the emails that were not based on the profile details in phase 1 and 2. It is observed that during phase 2 most of the emails were based on the victim's profile compared to phase 1.

Examining the tickler of the social media profile that led the trainees to choose the focus of their email, we compared the number of students focused on the occupation of the victim in both phases. It is observed that during phase 2, less emails were based on the victim's job than in phase 1. That is because in phase 2 trainees, based on the proposed methodology, discovered more interesting factors in the profile that can be used in a phishing email such as hobbies, likes, family, friends.

Another interesting factor was how the email addressed to the victim. To examine that aspect, we clustered the emails to personal (emails addressed just to the specific person using his name) and impersonal (with generic salutation). It is observed that during phase 2 more emails addressed the victim in person. This occurred because in phase 2 more emails than phase 1 were based on the detailed characteristics of the victim, for example, his food or training shoes preferences and not just the generic characteristic of him, his job as a pilot. In that sense, more trainees used a personal message as they created the email adjusted to him.

Focusing on the degree the message would be believable, we examined if there are any references about friends in the email message. It is observed that during phase 2 more emails included references to the victim's friends. That makes the message of the email more believable.

An example of trainees' phishing email can be seen in Figure 5.

An example of trainees' phishing email is the following:
Hello Thorsten,
I'm M.P. an aircraft spotter who always admired the activity of our "Luftwaffe". Don't worry :) Josh gave me your email because he liked a lot my idea and said me to forward it to you, so here you go. Both me and Josh know very well and I hope you too, that there are many pilots in our Air Force that lose their lives in every-day missions and I know that you are eager to sacrifice your own life for the homeland, which I really appreciate.
With the help of God my family and Joshes family begun a campaign in memory of our lost-in-action pilots. The goal is to organise a semi-marathon with an entry prise of 4€ each. The earnings will be given to the families of all your colleagues that lost their lives, as an action to soften their pain.
We have already a lot of people that confirmed their participation from all the branches of the armed forces and also civilians (!!!!).
Please, in memory of all them, bring your family and your friends and convince them to run, even for a little. It's for a good cause. You can find more info and also you can participate with this link http://.....

Glory to own wings
Best Regards
M.P.

P.S. Many greetings to Josh

Figure 5: An example of a spear phishing email that was created by a student

5. Conclusions

The base of our case study was a hypothetical person who was presented to the students from his social media profile. According to this profile, the students had to make two phishing emails in order to persuade him to click on a link. In phase 1, no guidance of how to prepare the email was presented to the participants, while in phase 2 the participants were asked to follow a methodology which guided them in how to think creatively and make a phishing email. Comparing the results of the two phases, it was detected that students following our proposed methodology made:

- phishing emails that were more elaborated, assessing the average words and the sentences per email
- more emails included detailed victim's personal information retrieved from his social media profile
- less emails were based only on the main characteristic of the victim, his job;
- more emails were personal, using his name and not a general greeting
- more emails had a reference to his friend

Considering the above, we believe that our proposed methodology gave the participants the chance to think as a real attacker who wants to achieve his/her goal. In other words, they didn't have to recognize or evaluate some emails and decide if these are suspicious as phishing but they had to construct a phishing email. Moreover,

the construction of their phishing email was not focused on the technical characteristics of it, but it was focused on the message included in it. Firstly, they had to look carefully at the profile of the victim and then decide how to make a believable message according to our suggested methodology. This is the main difference between our learning method and all the others that are already exist. Recognition of a phishing email from the technical view sometimes can be done even with the use of antivirus programmes. On the other hand, if everyone was aware that an attacker can make a believable message exploiting personal information available in public, then he/she will be more suspicious when receiving such an email.

Our learning methodology, however, had the limitation that it was applied on a rather unbalanced sample of cadets and students with the cadets outnumbering students and thus the sample may not be representative of the generic population. Moreover, the sample consisted of young people who may be more technologically savvy and therefore more familiar with emails, social media, etc. and thus our training methodology should be expanded in a wider audience consisted of a diverse group of people. In addition, our study could benefit from a qualitative comparison of the emails created in the two phases in order to reveal whether the emails created in the second phase are more believable than those created in the first phase. If that is the case, then the proposed anti-phishing training could be more effective since it enables participants to better understand the tactics of the spear phishing attackers.

References

Abbasi, A., Zahedi, F.M. and Chen, Y., 2016, September. Phishing susceptibility: The good, the bad, and the ugly. In *2016 IEEE conference on intelligence and security informatics (ISI)* (pp. 169-174). IEEE.

Arachchilage, N.A.G. and Cole, M., 2011, June. Design a mobile game for home computer users to prevent from "phishing attacks". In *International conference on information society (i-society 2011)* (pp. 485-489). IEEE.

Bonwell, C. C. and Eison, J. A., 1991. Active learning: Creating excitement in the classroom. Washington, *DC: Eric Clearinghouse on Higher Education*.

Branca (2023) 'Phishing Statistics - 2023', TRUELIST, 7 January. Available at: https://truelist.co/blog/phishing-statistics/ (Accessed: 07 January 2023).

Bruner, J. S., 1960. The Process of education. Cambridge, Mass.: Harvard University Press.

Canova, G., Volkamer, M., Bergmann, C. and Borza, R., 2014, September. NoPhish: an anti-phishing education app. In *International workshop on security and trust management* (pp. 188-192). Springer, Cham.

Clark, R., 2009. Accelerating expertise with scenario based learning. Learning Blueprint. Merrifield, VA: *American Society for Teaching and Development*.

Coulshed, V., 1993. Active learning: Implications forteaching in social work education. *British Journal of Social Work*, 23(1), 1-13.

Dodge, R., Coronges, K. and Rovira, E., 2012, June. Empirical benefits of training to phishing susceptibility. In *IFIP International Information Security Conference* (pp. 457-464). Springer, Berlin, Heidelberg.

Errington, E., 2005. Creating Learning Scenarios: A planning guide for adult educators. Palmerston North, NZ: CoolBooks.

Errington, E., 2010. Preparing Graduates for the Professions Using Scenario-based Learning. Australia: Post Pressed.

Estes, W. K., 1975. Handbook of learning and cognitive processes. Hillsdale, N.J., L. Erlbaurn Ass.

Evans, N., 1994. Experiential Learning for All. London, New York: Cassell

Felder, R., and Brent, R., 2003. Learning by doing. *Chemical Engineering Education*, 37(4), 282-283.

Gagne, R., 1975. Essentials of learning for instruction. New York: Dryden.

Golub, J., and NCTE Committee, 1988. Focus on Collaborative Learning: Classroom Practices in Teaching English. Urbana, IL; USA, National Council of Teachers of English Publishing.

Gupta, B.B., Tewari, A., Jain, A.K. and Agrawal, D.P., 2017. Fighting against phishing attacks: state of the art and future challenges. *Neural Computing and Applications*, 28(12), pp.3629-3654.

Jagatic, T.N., Johnson, N.A., Jakobsson, M. and Menczer, F., 2007. Social phishing. *Communications of the ACM*, *50*(10), pp.94-100.

Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L.F., Hong, J. and Nunge, E., 2007, April. Protecting people from phishing: the design and evaluation of an embedded training email system. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 905-914).

Margetson, D., 1998. What counts as problem-based learning? *Education for Health*. 11:193-201.

McCurdy, R. (2022) 'The Biggest Phishing Breaches of 2022 and How to Avoid them for 2023', *Security Boulevard*. Available at: https://securityboulevard.com/2022/11/the-biggest-phishing-breaches-of-2022-and-how-to-avoid-them-for-2023/ (Accessed: 07 January 2023).

Meyers, C., and Jones, T. B., 1993. Promoting active learning. San Francisco, CA: Jossey-Bass.

Sumner, A., Yuan, X., Anwar, M. and McBride, M., 2021. Examining factors impacting the effectiveness of anti-phishing trainings. *Journal of Computer Information Systems*, pp.1-23.

Pars, C., 2017. *PHREE of Phish: the effect of anti-phishing training on the ability of users to identify phishing emails* (Master's thesis, University of Twente).

Roepke, R., Koehler, K., Drury, V., Schroeder, U., Wolf, M.R. and Meyer, U., 2020, September. A pond full of phishing games-analysis of learning games for anti-phishing education. In *International Workshop on Model-Driven Simulation and Training Environments for Cybersecurity* (pp. 41-60). Springer, Cham.

Roepke, R., Drury, V., Meyer, U. and Schroeder, U., 2022. Exploring and evaluating different game mechanics for anti-phishing learning games. *International Journal of Serious Games*, 9(3), pp.23-41.

Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L.F., Hong, J. and Nunge, E., 2007, July. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. *In Proceedings of the 3rd symposium on Usable privacy and security* (pp. 88-99).

Thomsen, B. C., C. C. Renaud, S. J. Savory, E. J. Romans, O. Mitrofanov, M. Rio, S. E. Day, A. J. Kenyon, and J. E. Mitchell. 2010. "Introducing Scenario Based Learning: Experiences from an Undergraduate Electronic and Electrical Engineering Course." Paper presented at *IEEE global engineering education conference (EDUCON)*, Madrid, April.

Vayansky, I. and Kumar, S., 2018. Phishing–challenges and solutions. *Computer Fraud & Security*, *2018*(1), pp.15-20.

Vygotsky, L., 1978. Interaction between Learning and Development. In Gauvain & Cole (Eds.) *Readings on the Development of children*. New York: Scientific American Books. pp. 34-40.

Watson, J. B., 1930. Behaviorism (Revised edition). Chicago: University of Chicago Press.

Wen, Z.A., Lin, Z., Chen, R. and Andersen, E., 2019, May. What. hack: engaging anti-phishing training through a role-playing phishing simulation game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).

Designing Security for the Sixth Generation: About Necessity, Concepts and Opportunities

Christoph Lipps[1], Annika Tjabben[1], Matthias Rüb[1], Jan Herbst[1], Rekha Reddy[1], Sogo Pierre Sanon[1],Yorman Munoz[1] and Hans Dieter Schotten[1,2]

[1]Intelligent Networks Research Group, German Research Center for Artificial Intelligence, Kaiserslautern, Germany

[2]Institute for Wireless Communication and Navigation, RPTU Kaiserslautern-Landau, Kaiserslautern, Germany

Christoph.Lipps@dfki.de
Annika.Tjabben@dfki.de
Matthias.Rueb@dfki.de
Jan.Herbst@dfki.de
Sogo_Pierre.Sanon@dfki.de
Rekha.Reddy@dfki.de
Yorman.Munoz@dfki.de
Hans_Dieter.Schotten@dfki.de

Abstract: Intelligent, comprehensive and, above all, secure wireless interconnection is the driving force behind technological progress. To ensure this, the development towards Sixth Generation (6G) Wireless Systems has been launched and is scheduled to be operational by 2030. This data technology of the future turns 6G into the infrastructure of a new generation of mobile, intelligent, and context-sensitive services, available everywhere and featuring high trustworthiness and performance, relying on both, network-side and off-network context sources. In addition, the networks themselves ought to become intelligent and thus more efficient and resource-saving, which requires a high degree of automated utilization of Artificial Intelligence (AI). Building upon the principles of information and communication theory for both the physical (bit)-transmission layer (PHY) and media access, new communication concepts for 6G will be developed providing the foundations for research into new single and multi-user operation, access and core networks.

The flip side of this coin of opportunities: Sophisticated technology inevitably leads to additional security vulnerabilities, open access systems and Open-Radio Access Network (O-RAN) approaches imply new attack vectors. The holistic interconnection of everything renders it ever more attractive to attackers to harm systems, and create damage. Furthermore, enhanced computational power along with quantum computers make conventional systems more vulnerable than ever, and the value of the transmitted data increases tremendously: It is not only machine and sensor data, but also very personal and healthcare data transmitted with 6G.

Therefore, the aim is to build a resilient and secure 6G system capable of recognizing attacks and uncertainties, flexibly absorbing them, recovering in a timely and sustainable manner, and compensating for impaired functionality through transformation. This holistic resilience-by-design approach is based, among other things, on technology such as Quantum Key Distribution (QKD) and Post Quantum-Crypto to achieve end-to-end security, Reconfigurable Intelligent Surfaces (RISs) to rely, control and manipulate the wireless transmission channel, Wireless Optical Communication (WOC), Physical Layer Security (PhySec), but also Body Area Networks (BANs), the integration of the human body relying on biometrics and the Tactile Internet (TI). These concepts will be discussed and shed light on in the scope of this work.

Keywords: Physical Layer Security; Physically Unclonable Functions; Human-PUFs; Channel-PUFs; (Cyber)Security; Sixth Generation; Body Area Networks (BANs);

1. The Next Generation Mobile Network: The Sixth Generation

The level of interconnection is continuing apace: While wired connectivity has been the standard in many areas, ranging from private applications to complex industrial networks, just a few years ago, a trend towards wireless connectivity is rapidly developing. Above all, the advantages of scalability (no need for cumbersome new cables as components are added), flexibility (no restriction on location), and cost efficiency (connectors and cables are expensive) speed up the process even more. And even if there is already a broad portfolio of technologies and standards, like Long Range Wide Area Networks (LoRaWAN), ZigBee, Industrial Radio and IEEE 802.11ax Wireless LAN -with the push to the 6GHz frequency range-, each associated with different advantages and disadvantages, (cellular) mobile radio is shifting into the scientific spotlight: Although the Fifth Generation (5G) Wireless System is currently in the global roll-out phase there are strong arguments for already considering its successor. The International Telecommunication Union (ITU), for instance, estimated that the total number of Mobile Broadband (MBB) subscribers worldwide will reach 17.01 billion by 2023 and Machine-to-Machine (M2M)

terminals will cross the 100 billion threshold within this time frame (International Telecommunication Union, 2015). Alongside the mere number of devices, the requirements for the technology are rising as well: disruptive innovations such as Holographic-Type Communications (HTC), Augmented- and Virtual Reality (AR/VR) towards an (industrial) metaverse, (Medical) Digital Twins ((M)DTs), Tactile Internet (TI) and a ubiquitously global connectivity are (over)demanding the existing infrastructure. Nevertheless, there is currently no single technology which can meet the whole range of requirements. Instead, contrary to this, and among other things, this is the key aspect during the development of 6G. Different technologies are considered as "6G enabling technologies": heterogeneous Radio Access Technology (RAT) and Open – Radio Access Networks (O-RANs) should provide flexibility, virtualization and intelligence while defining disaggregation and openness; the available and feasible spectrum will include mmWave and (sub)Terahertz (THz) ranges, including the spectrum of visible light as well as the dynamic management of the spectrum; Wireless Optical Communication (WOC) and Visible Light Communication (VLC) shall enable applications such as "data kiosks" and indoor communication unaffected by Radio Frequency (RF) interference; Reconfigurable Intelligent Surfaces (RISs), Intelligent Reflecting Surfaces (IRSs) and meta materials in general, shall be capable of controlling the wireless channel and its properties and of steering and relaying electromagnetic waves in the sense of sustainability and GreenICT as passive but controllable elements; Joint Communication And Sensing (JCAS) will enable a symbiosis between mobile communication and mobile sensing, allowing robots and Extended Reality (XR) applications to scan and perceive their environment (in indoor and outdoor environments); methods of Artificial Intelligence (AI) will have a huge impact, not simply because a core concept will be distributed computing, processing, and putting AI in the edge, but also because AI will be integrated into the air interfaces and applied as the basis for signal detection, synchronisation, channel estimation, equalisation, as well as modulation and coding. And all of these technological enablers, whose interconnections are highlighted in **Figure 1**, will have a direct impact on the security of the systems, the selection of appropriate security applications and the implementation of security measures.

The overall necessity for secure applications, as well as a general perspective, is described in Section 2. Based on this, the enabling technologies mentioned are specified in more detail in Section 3. These are evaluated, discussed and set in relationship to each other in Section 4, afterwards, Section 5 concludes the work and gives an outlook.

Figure 1: 6G as an enabler for the interconnection of the future of everyday life. In addition to increased availability, one of the design goals is to integrate "intelligence" into the edge and to incorporate the methods of artificial intelligence more effectively. This is achieved through a set of enabling technologies, including Wireless Optical Communication, Reconfigurable Intelligent Surfaces, Ultra-Massive MIMO and the usage of higher frequency ranges compared to the past.

2. Security of the Sixth Generation: A General Perspective

The demand for security and reliability is certainly as old as humanity itself: Potential threats emanating from strangers require detection and appropriate assessment; Trust has to be established, maintained and extended over time; Access and disclosure of information to untrustworthy persons shall be restricted. This is nothing particularly new, however, with increasing globalisation, the interconnection of everything with each other and the possibility of remote access from anywhere, the need for technological is increasing further. In addition, the development of the Sixth Generation Wireless Systems is oriented towards the infrastructure: the focus should be on people, and services and "intelligence" should be decentralized and brought to the edge and to the applications. This is accompanied by a particular need for security and privacy, especially if sensitive and personal information is transferred.

Already during the development of the previous version, the Fifth Generation (5G) mobile radio, self-optimisation, self-healing and self-configuration of the network had been one of the design criteria. This aspect is taken into account and extended towards a resilient infrastructure, which is covering the holistic cyber-resilience lifecycle including the detection of anomalies and threats, appropriate deception technology -to mislead, distract and gather (attack) information from the attacker-, but also mitigation and defensive applications as well as active encryption and recovery methods.

In addition, methods of Artificial Intelligence will be integrated into all parts of the processing stack. As AI is another enabling technology, and of pervasive and of pivotal relevance, trusted AI (Cohen, et al., 2019), trustworthy AI (Kaur, et al., 2022) and Machine Learning (ML) (Porambage, et al., 2021) addressing the privacy and security of the systems (Ma, et al., 2022), will be a crucial issue, also in terms of the acceptance of a wide range of people in the use of 6G and AI in general. The decentralisation of resources and the associated approach of integrating intelligence into the edge is achieved by methods of federated learning, secured, for instance, by Fully Homomorphic Encryption (FHE) (Sanon, et al., 2023).

Based on Industrial Internet of Things (IIoT) approaches to achieve integrity and authenticity of the entities involved (Lipps, et al., 2019), Trusted Execution Environment (TEE), Trusted Platform Modules (TPMs), Physically Unclonable Functions (PUFs) and flexible trust anchors (Ziegler, et al., 2021) will be taken into account to harden the multi-stakeholder environments against the multitude of potential attacker scenarios and -vectors.

Besides the general concepts, there are a number of key enabling technologies which in turn can have an individual impact on the security of the system.

3. Enabling the Sixth Generation: Technologies and their Security Implications

Starting in 2020, the development of Sixth Generation mobile communications is currently gaining momentum in Europe (Jiang, et al., 2021). Thereby, a decisive design feature is the integration and diversification of different technological approaches, so-called key enabling technologies: Due to the variety of requirements placed on the future infrastructure, there is no one-size-fits-all solution, but rather a conglomerate of these technologies deliberately intended to operate together and interact with each other.

3.1 (Sub) Terahertz Communication

One aspect of the 6G discussion is the use of higher frequency ranges in the Terahertz spectrum from 95 GHz up to 10 THz (Singh & Sicker, 2020), for which there are a number of relevant reasons to take a closer look: For numerous use cases such as Wireless Sensor Networks & Wireless Body Area Networks (BANs), (Unmanned) Aerial Vehicles ((U)Avs), Intelligent Transport Systems (ITS), Holographic-Type Communications (HTC) and application in the medical sector, this frequency range is considered particularly advantageous (Lipps, et al., 2021). One reason therefore is the reduced interference with the common spectrum which would unload the data traffic of these frequencies (Singh & Sicker, 2020). Furthermore, due to the huge potentially applicable frequency band, it is considered as an enabler for Ultra-high throughput applications. Applications like Tactile Internet, Extremely Dense Networks (EDN)s & Internet of Things (IoT) will benefit therefrom.

However, the shift to higher frequencies is also of relevance particularly in terms of security: Vulnerabilities of commonly used spectrum related to Fake Base Station-, Man-in-the Middle (MitM)-, and Denial of Service (DoS) attacks, as well as user impersonation and eavesdropping could be remedied significantly. A lower coverage and lower penetration power renders the system inherently more secure, as, for instance, Non-Line-of-Sight (N-LoS) attacks can be blocked and mitigated by performing enhanced beamforming and smart shielding. The higher

energy radiation -in comparison to the common spectrum-, has a lower penetration depth into different materials which leads to fast absorption in N-LoS situations. Besides, systems working in the scope of anti-jamming could be enhanced: Due to the large frequency range of the THz-Spectrum, the approach of Frequency Hopping (FH) for anti-jamming could be extended. The aspect of enhanced beamforming adds an extra level of security against jammers; attackers would have to perfectly align to the signal which is a almost impossible task, due to the complex technology needed to achieve this in an undetected manner (Singh & Sicker, 2020).

3.2 Reconfigurable Intelligent Surfaces

Reconfigurable Intelligent Surfaces are another emerging topic in the context of the new generation wireless communications. Building up of several, commonly passive, antennas elements these “man-made surfaces” (Pan, et al., 2021) are capable of manipulating the reflecting characteristics of electromagnetic waves. To achieve this, such surfaces consist of a multitude of sub-wavelength wide elements each of which independently controllable and adjustable. Specific modification of the characteristic enables an impinging wave signal to be reflected in the desired direction.

These attributed features of being able to manipulate a channel have a direct influence on security, therefore *Lipps et al.* (2022) specified conceivable scenarios in which RIS can be used to increase this security: i) RIS-enabled beamforming and -splitting, which can mitigate the risk of eavesdropping through the possibility of steered communication towards legitimate users. Besides, tighter beams up to pencil-sharp beamforming can chance the overall properties of a radio link from a broadcast medium to a targeted communication. ii) RIS-based Anti-jamming, whereby the mentioned broadcast characteristic renders the system in general vulnerable to conscious and unconscious interferences, and RISs can help by destructive and constructive superimposition of signals, as well as by phase-shifting the reflected signals; iii) RIS-manipulated channel-profiles, which are the basis for Physical Layer Security (PhySec) Secret Key Generation (SKG) and whose key derivation can be enhanced by additionally introduced entropy, thereby increasing security; iv) RIS-individual fingerprints, where, based on the principle of Physically Unclonable Functions (PUFs). It can be assumed that each RIS exhibits a hardware-based individual fingerprint and thus characteristic properties which can be measured and included as a security feature, and v) RIS-supported additional context information, which, similar to unintentional fingerprinting, can also be actively and consciously used as an additional context.

Besides, the combination of RISs with other technologies such as VLC is a conceivable approach to enable the combined potential of both. *Rüb et al*. (2022) are discussing some of the benefits regarding the use of laser-based VLC with RIS and are describing approaches towards the security of a VLC system.

3.3 Wireless Optical Communications

With increasing interconnectivity there is a rising number of devices communicating with each other. As the available Radio Frequencies spectrum is limited, this will cause major challenges in the future. Besides the shift of higher THz frequency ranges, the consideration of Optical Wireless Communication is discussed.

This form of communication uses the broad unregulated optical spectrum, which is sub-divided into Infrared, Visible Light and Ultraviolet, each giving rise to different implementations for high-speed wireless communication. Furthermore, for OWC, in contrast to RF, the interference problem does not arise. Current research on OWC focuses on establishing a connection with a sufficiently high data rate (an overview of recently achieved data rates is given in (Elgala, et al., 2011)) while at the same time ensuring the security of the transmission.

At first glance OWC seems to provide a higher level of security due to the transmission being limited by walls and obstacles. Yet the low cost and limited computing power of many devices renders them unsuitable for standard cryptographic approaches (Anon., 2019). Since OWC differs fundamentally from other wireless networks different approaches need to be followed to ensure secure transmission. One method tries to adapt different modulation schemes to bring a positive benefit to the security of the system. Recently *Panayirci et al.* (2020) apply the Spatial modulation technique, which reduces inter-channel interference, together with a novel spatial constellation design technique aimed at enhancing the physical layer security. Furthermore, it turns out that OWC is vulnerable to integrity attacks. *Soderi and De Nicola* (2021) show that the Watermark Blind Physical Layer Security (WBPLSec) algorithm, a combination of watermarking and jamming, can be used to ensure secure communication against these attacks. These recent approaches are promising, but they only consider single security aspects. To find a general solution for the security of OWC further research is required.

3.4 Localisation & Sensing

Robots and (U)AVs are just one aspect of future technologies expected to be a part of our everyday life. As by now autonomous driving vehicles can already be seen on the streets. Mobile working robots and vehicles should record their 3D surrounding, which mostly is done using radar, camera, or spectroscopy systems. To meet the demand for accuracy in sensing and supported data rates of up to 100 Gb/s the idea is to combine both, communication and sensing together into a joint system and frequency spectrum (Fettweis et. al., 2021). The mobile spectrum itself will barely support needed data rates with the common frequency spectrum, on the other hand, mobile resources for broadband communication are only rarely needed at the same time for mobile sensing. JCA&S is the integration of both, communication and localization sensor technology in one common system. *Fang et al.* (2022) pointed out, that regarding security there is a conflict of interest between those topics as the sensing part needs to interact strongly with the environment which increases the risk of eavesdropping on the communication. Additionally, they highlight the risk of sensing targets being malicious, as in contrast to communication they are not checked for legitimacy.

Another issue is the information leakage that could occur between the two activities. *Günlü et al.* (2022) investigate several modelling approaches for the complex interplay of sensing performance and secrecy performance which are both evaluated based on the signal strength at the sensed target, which could be malicious. In general, a convincing security solution to address the above-mentioned issues requires more future investigations.

3.5 Link Integrity Monitoring and Anomaly Detection

Integrity monitoring comprises the regular review of available information regarding correctness and origin. This is an important concept in a variety of applications, like position parameters of vehicles in the scope of self-driving cars and sensor data in IoT networks, for example, in agriculture scenarios. Link integrity monitoring of wireless communication refers to the question if an established link is originated by a legitimate source or a malicious third party. For upcoming wireless networks, it is necessary to evaluate, if it could be possible to transfer existing concepts of integrity monitoring with methods of AI onto the specific use-case of link integrity monitoring in wireless networks.

Furthermore, integrity monitoring is closely related to anomaly detection, as non-legitimate sources will stand out for example in their high-level parameters (data rate, time of connection, reception angle) or the content of information. With the dynamic adaption of attackers and the expected high density of IoT devices, state-of-the-art solutions for Intrusion Detection Systems (IDS) are no longer sufficient (Oleiwi, et al., 2022). Recent works have applied new combinations of different AI approaches including data preparation and feature extraction with classic machine learning algorithms (Mittal, et al., 2021) and the final step of intrusion detection with deep learning (Gupta, et al., 2022). Especially promising is the combination of several models in an ensemble learning approach to detect anomalies (Jaw & Wang, 2021). *Oleiwi et al.* (2022) proposed in 2022 an approach for intrusion detection with two modified classic algorithms: random forest and support vector machine, achieving an accuracy above 99 %.

Recent approaches especially focus on energy efficiency, as multiple AI algorithms for intrusion detection – especially deep learning techniques– require a high computing power. Especially for dynamic systems which require retraining of existing artificial neural networks could be very energy consuming. In summary, integrity monitoring and anomaly detection in upcoming wireless networks will most likely consist of ensemble models of multiple machine learning approaches. One major concern will be to make the utilized algorithms energy efficient to positively impact the trade-off between cost and security.

3.6 Security in the Post-Quantum era

Post-Quantum Cryptography (PQC) constructs and studies cryptosystems that are secure against both quantum and classical computers. Existing public-key cryptosystems are based on the difficulty of factoring and the difficulty of calculating elliptic curve discrete logarithms, but then, *Shor (1994)* developed a quantum algorithm which can effectively solve those problems. Hence, proving that quantum computers would eventually be able to break current security systems in a relatively short amount of time.

In recent years, many systems conjectured to resist quantum attacks have been proposed for the replacement of current public-key protocols. This prompted the National Institute of Standards and Technology (NIST) to start a standardization process (Chen, et al., 2016), to produce strong quantum cryptographic standards which are

expected to appear in 2024. NIST is well known for standardizing cryptographic algorithms. For instance, the symmetric-key algorithm Advanced Encryption Standard (AES), which is widely used nowadays, was established by NIST in 2001 after a process like the PQC Standardization (Nechvatal, et al., 2001). The PQC algorithms established by NIST are expected to be adopted in most use cases around the world.

At the time of writing, NIST has already completed three rounds in the selection process and has already selected four algorithms to be standardized, one public-key encryption algorithm and three digital signature algorithms. There are four additional algorithms, all for public-key encryption, that will go through the fourth round and any of them can be standardized. More details on the algorithms can be found in (Alagic, et al., 2022).

As wireless communication becomes increasingly important and use cases getting diverse more than ever, the security aspect becomes crucial. PQC needs to be considered in future generations and incorporated in their architectures. In addition, further investigations regarding the most suitable algorithm for wireless communication among those that will be standardized need to be conducted.

4. Discussion and Evaluation of 6G-enabling Technologies

The development of the Sixth Generation Wireless Systems will not rely on a specific technology, but rather on different aspects, each associated with individual strengths and weaknesses. Yet, as already indicated, they cover a wide range of topics, from Reconfigurable Intelligent Surfaces to Wireless Optical Communication to Post-Quantum Security. To still provide a statement about the suitability and usability of the applications in possible uses-cases, several Key Performance Indicators (KPIs) are to be defined, on the basis of which an evaluation can be carried out. However, these KPIs only represent a high-level overview and have to be fine-tuned to be finer-grained and more specific in further work:

- **Detection**: This refers to an assessment of the extent to which the technology can contribute to the detection of attackers and intruders.
- **Prevention**: The ability to stop intruders from stealing or altering information of the legitimate communication link.
- **Cost**: Total cost of a system including the initial setup/purchase and continuous cost like energy consumption.
- **Applicability**: For a widespread application, it is necessary that the system does not affect the performance and functionality of the communication link.
- **Usability:** Describes how practicable a technology is (especially in daily use).
- **State of development:** This KPI evaluates the subjective state of research regarding the topic. Especially many open challenges will result in a low evaluation.

4.1 (Sub)THz communication

The usage of higher frequency ranges offers advantages, especially with regard to the possible data rate and the applicability in combination with localisation, but also entails disadvantages such as higher susceptibility to interference. Nevertheless, it remains a broadcast system rendering it difficult to detect attackers and intruders. However, tighter beams allow for narrower band communication and can make eavesdropping more difficult. Another limiting factor can be the communication itself: The usage of higher frequencies is associated with high path losses and low coverage in the range of only a few meters. Thereby, this limits the accessibility, does not allow long-range communication and indicates at least the need for expensive communication hardware. In addition, various materials and molecules have their absorption resonance in the range of this frequency space, leading to high spreading losses. Even for short communication distances between the THz links sensitive user information are needed and must be shared across the wireless system, for instance, user´s localization, device orientation, mobile patterns, and surrounding environments up to blockages because of LOS requirements.

(Singh & Sicker, 2020) therefore talked about the “Gap of THz” communication and the lack of efficient and cheap THz signal generators as well as receivers. They propose to use existing and developed technologies in the field of imaging/ spectroscopy, which are having increased knowledge, hardware, and detectors related to THz.

4.2 Reconfigurable Intelligent Surfaces

RISs are a frequently discussed and active field of research. Even though the idea is not completely new, the approach provides advantages, especially when used as a passive element and in terms of sustainability and

GreenICT. These increase the flexibility of the communication system and allow the radio channel to be tailored to the current conditions and needs. However, if the RIS implementations are used in combination with higher frequencies (THz) range, they are offering “raw-power” in terms of transmission rate, but are very inflexible due to the limited transmission range due to high absorption not only by air but also by (moving) obstacles like persons or vehicles. In addition, it is not yet clear how large functional RIS will really be in the end. Their acquisition and construction - especially using PIN diodes - is still very expensive and there are few to no commercial manufacturers on the market. What remains to be observed is the effect once the corresponding meta-materials are available and can be used accordingly.

4.3 Wireless Optical Communication

The approach of using (visible) light for communication is also not entirely new; this technology has been proven and is effective, especially in fibre-optic and photonic systems. Using wireless technology - which is likewise just electromagnetic waves - poses challenges, however, as systems are interfered with by natural light, sunlight and lamps, for example. On the other hand, the systems are not affected by RF signals, which makes a combination with them attractive. For security applications: visible light is already blocked by simple walls, in contrast to RF signals, which complicates eavesdropping and at the same time increases the probability of detecting an attacker. Moreover, the technology and the use of for instance lasers are well researched and open source systems already exist for the utilization and simple implementation.

4.4 Localisation & Sensing

Joint Communication and Sensing offers promising possibilities, particularly in combination with high frequencies in the THz range and the associated accuracy in the sub-centimetre range, although the above-mentioned effects and disadvantages of THz must be taken into account. In combination with the possibility of tight beams and beam-steering, JCAS offers the opportunity of targeted communication and thereby minimizing jamming, spoofing and eavesdropping, as the directions of the attackers can be localized and avoided. In addition, it is a technology that requires either no or very few additional hardware, but rather is to be integrated into existing communication systems, making it a by-product. This reduces costs, which in turn makes it more attractive for industrial applications.

4.5 Link Integrity Monitoring and Anomaly Detection

Monitoring and verifying the integrity of network nodes in operation is an essential component of network security and cyber resilience. By monitoring anomalies, suspicious events in networks are detected and appropriate countermeasures such as re-routing, shifting of resources or shutdown are made possible. By integrating methods of artificial intelligence, which only require appropriate computing power, these systems can be operated at rather low costs, but provide important and essential information about the status of a network. With appropriate configuration, many of these systems are quite applicable, but can be strongly enhanced with appropriate expert knowledge.

Table 1: Key performance indicators for the enabling Technologies discussed, related to security applications. (The evaluation of the different approaches is indicated as positively (+) / average (o) / below average (-). If current research is not feasible yet to give a convincing indication, the topic is marked as under investigation (UI).

	(sub)THz communication	RIS	WOC	Localization & Sensing	Link Integrity Monitoring & Anomaly Detection	Post-Quantum Security
Detection	o	+	+	+	+	o
Prevention	+	+	o	-	+	+
Cost	o	UI	+	+	+	-
Applicability	+	+	+	+	+	+
Usability	+	o	+	+	+	o
State of Develop-ment	+	-	+	UI	o	-

4.6 Post-Quantum Security

Developments in quantum communications are a sword of Damocles hanging over the current security landscape. Experts often postulate that breakthroughs in quantum technology will have a massive impact on existing cryptographic methods, as they can be broken in a fraction of the time it used to take. However, these developments i) are currently not available and ii) there are also ambitions for quantum security applications such as quantum tokens. Nevertheless, it is certainly sensible to rely on quantum safe solutions such as Physical Layer Security (at least as complement/extension and additional factor), as these are not based on computational complexity but rather on inherent secrets and characteristics.

5. Conclusion and Future Work

The work discusses the ideas and directions emerging in the current development of the Sixth Generation Wireless Systems, being worked on in different research directions. Thereby, starting from expectations of applications such as Holographic-Type Communication, Tactile Internet, Augmented-/Virtual- and Extended Reality, (Medical) Digital Twins and the (industrial) metaverse in general, technologies have been identified in order to satisfy them. Because these applications not only transfer much more business-sensitive information, but also personal data and data requiring special protection. The security requirements are perhaps higher than ever before. For this purpose, the 6G-enabling technologies were not just described in this work, but their influence and impact on security and security applications have been discussed as well.

All of the technologies discussed have a direct impact on security, but are at different stages of development. Reconfigurable Intelligent Surfaces, for instance, are a frequently discussed topic within the research community (even if the idea is not that new) and a lot of time and resources are being invested right now. Although these are rather expensive to produce (particularly the PIN diodes), and it is currently assumed that larger surfaces will be required (in the direction of the walls of the houses), which remains open for discussion. However, since a lot of research is being done in this field right now, it is worthwhile to focus on security requirements as well and to consider these fundamental aspects. Generally speaking, this is the major advantage of the current stage of development: 6G is still in the development phase, with the first steps toward standardization not being taken before roughly 2025. Right now, fundamental aspects such as Security-as-a-Service, Security-by-Design and the trend towards Post-Quantum Security can be taken into consideration within the development. It is essential to raise awareness and to already create broad acceptance, not only among the scientific community, but also among future applicants, not only for technology 6G, but also and above all for the methods of Artificial Intelligence. After all, this is where a significant amount of work will be required and where is potential for further research: trusted AI and trustworthy AI.

Acknowledgment

This work has been supported by the Federal Ministry of Education and Research of the Federal Republic of Germany (Förderkennzeichen 16KIS1283 AI-NET PROTECT). The authors alone are responsible for the content of the paper.

References

Alagic, G., Apon, D., Cooper, D., Dang, Q., Dang, T., Kelsey, J., Lichtinger, J., Miller, C., Moody, D., Peralta, R., Perlner, R., Robinson, A., Smith-Tone, D. and Liu, Y.-K., "Status Report on the Third Round of the NIST Post-Quantum Cryptography Standardization Process", *Information Technology Laboratory, Computer Security Resource Center,* National Institute of Standards and Technology, Gaithersburg, MD, USA, 2022.

Chen, L., Jordan, S., Liu, Y.-K., Moody, D., Peralta, R., Perlner, R. and Smith-Tone, D., "Report on Post-Quantum Cryptography", *Computer Security Division, Applied and Computational Mathematics Devision, Information Technoloy Laboratory,* National Institute of Standards and Technology, Gaithersburg, MD, USA, 2016.

Chorti, A., Barreto, A.N., Köpsell, S., Zoli, M., Chafli, M., Sehier, P., Fettweis, G. and Poor H.V., "Context-Aware Security for 6G Wireless: The Role of Physical Layer Security", *IEEE Communications Standards Magazine,* vol. 6, no. 1, pp. 102--108, DOI: 10/1109/MCOMSTD.0001.2000082, 2022.

Cohen, R., Schaekermann, M., Liu, A. and Cormier, M., "Trusted AI and the Contribution of Trusted Modelling in Multiagent Systems", *Proceedings of the 18th International Conference on Autonomous Agents MultiAgent Systems,* pp. 1644--1648, DOI:1.5555/3306127.3331890, 2019.

Elgala, H., Mesleh, R. and Haas H., "Indoor Optical Wireless Communication: Potential and State-of-the-Art", *IEEE Communications Magazine*, vol. 49, no. 9, pp:56--62, DOI:10.1109/MCOM.2011.6011734, 22011.

Fang, X., Feng, W., Chen, Y., Ge, N. and Zhang, Y., "Joint Communication and Sensing: Models and Potentials of Using MIMO", *ArXiV - Electrical Engineering and Systems Science - Signal Processing*, arXiv:2205.09409, 2022.

Günlü, O., Bloch, M., Schaefer, R.F. and Yener, A., "Secure Joint Communication and Sensing", *IEEE International Symposium on Information Theory (ISIT)*, Espoo, Finland, DOI: 10.1109/ISIT50566.2022.9834748, 2022.

Gupta, N., Jindal, V. and Bedi, P., "CSE-IDS: Using cost-sensitive deep learning and ensemble algorithms to handle class imbalance in network.based intrusion detection systems", *Computer & Security*, vol. 112, DOI: 10.1016/j.cose.2021.102499, 2022.

IEEE Standard for local and Metropolitan Area Networks -- Part 15.7: Short-Range Wireless Optical Communication Using Visible Light, IEEE Std 802.15.7-2011, DOI:10.1109/IEEESTD.2011.6016195

International Telecommunication Union, "IMT Traffic estimated for the years 2020 to 2030", M Series Mobile, radiodetermination, amateur and related satelliteservices Report ITU-R M.2370-0.

Jaw., E. and Wang, X., "Deature Selection and Ensemble-Based Intrusion Detection System: An Efficient and Comprehensive Approach", *symmetry,* vol. 13, no. 10, DOI:10.3390*sym13101764, 2021.

Jiang, W., Han, B., Habibi, M.A. and Schotten, H.D., "The Road Towards 6G: A Comprehensive Survey", *IEEE Open Journal of the Communications Society*, vol. 2, pp. 334--366, DOI:10.1109/OJCOMS.2021.3057679, 2021.

Kaur, D., Uslu, S., Rittichier, K.J. and Durresi, A., "Trustworthy Artificial Intelligence: A Review", *ACM Computing Surveys,* vol. 55, no. 39, pp. 1--38, DOI: 10.1145/3491209, 2022.

Lipps, C., 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; 25th ITG-Symposium*, Osnabrück, Germany, 2021.

Lipps, C., Duque Anton, S. and Schotten, H.D., "Enabling Trust in IIoT: A PhySec Based Approach", *International Conference on Cyber Warfare and Security*, Stellenosch, South Africa, 2019.

Lipps, C., Herbst, J., Reddy, R., Franke, L., Becker, S., Rahm, M. and Schotten, H.D., "Reconfigurable Intelligent Surfaces: A Physical Layer Security Perspective", *4th International Conference on Data Intelligence and Security (ICDIS)*, Shenzhen, China, DOI: 10.1109/ICDIS55630.2022.00034, 2022.

Lipps, C., Weinand, A., Krummacker, D., Fischer, C. and Schotten, H.D., "Proof of Concept for IoT Device Authentication Based on SRAM PUFs Using ATMEGA 2560-MCU", *1st International Conference on Data Intelligence and Security (ICDIS)*, South Padre Island, TX, USA, DOI: 10.1109/ICDIS.2018.0001, 2018.

Ma, C., Li, J., Wei, K., Liu, B., Ding, M., Yuan, L., Han, Z. and Poor, V., "Trusted AI in Multi-Agent Systems: An Overview of Privacy and Security for Distributed Learning", *Computer Science - Distributed, Parallel, and Cluster Computing*, arXiv:2202.09027, 2022.

Mittal, M., de Prado, R.P., Kawai, Y., Nakajima, S. and Munoz-Exposito, J.E., "Machine Learning Techniques for Energy Efficient and Anomaly Detection in Hybrid Wireless ensor Networks", *energies*, DOI: 10.3390/en14113125, 2021.

Nechvatal, J.R., Barker, E.B., Bassham, L.E., Burr, W.E., Dworkin, M.J., Foti, J. and Roback, R., "Report on the Development of the Advanced Encryption Standard (AES)", *Journal of Research (NIST JRES),* National Institute of Standards and Technology, Gaithersburg, MD, USA, 2001.

Oleiwi, H., Mhawi, D.N. and Al-Raweshidy, H., "MLTs-ADCNs: Machine Learning Techniques for Anomaly Detection in Communication Networks", *IEEE Access*, vol. 10, pp. 91006--91017, DOI: 10.1109/ACCESS.2022.3201869, 2022.

Panayirci, E., Yesilkaya, A., Cogalan, T., Poor, H.V. and Haas, H., "Physical-Layer Security With Optical Generalized Space Shift Keying", *IEEE Transactions on Communications,* vol. 68, no. 5, pp. 3042--3056, DOI: 10.1109/TCOMM.2020.2969867, 2020.

Pan, C., Ren, H., Wang, K., Kolb, J.F., Elkashlan, M., Chen, M., Di Renzo, M., Hao, Y., Wang, J., Swindlehurst, A.L., Yoi, X., and Hanzo, L., "Reconfigurable Intelligent Surfaces for 6G Systems: Principles, Applications, and Research Directions", *IEEE Communications Magazine*, vol. 59, no. 6, DOI: 10.1109/MCOM.001.2001076, 2021.

Porambage, P., Gür, G..m Osorio, D.P.M, Livanage, M. and Yilanttila, M., "6G Security Challenges and Potential Solutions", *Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit),* Porto, Portugal, DOI: 10.1109/EuCNC76GSummit51104.2021.9482609, 2021

Rüb, M., Tjabben, A., Munoz, Y., Grüber, J., Ahr, P., Herbst, J. and Lipps, C., "A Symbiotic 6G Enabler: Combining Laser based Visible Light Communication with Reconfigurable Intelligent Surfaces", *Workshop on Next Generation Networks and Applications (NGNA),* Kaiserslautern, Germany, DOI: 10.13140/RG.2.2.24818.38084, 2022.

Sanon, S. P., Reddy, R., Lipps, C. and Schotten, H.D., "Secure Federated Learning: An Evaluation of Homomorphic Encrypted Network Traffic Prediction", *IEEE Consumer Communications & Networking Conference*, Las Vegas, USA, 2023.

Shor, P.W., "Algorithms for Quantum Computation: Discrete Logarithms and Factoring", *Proceedings 35th Annual Symposium on Foundations of Computer Science*, Santa Fe, NM, USA, DOI: 10.1109/SFCS.1994.365700, 1994.

Zeng, S., Zhang, H., Di, B., Han, Z. and Song, L., "Reconfigurable Intelligent Surfaces /RIS) Assisted Wireless Coverage Extension: RIS Orientation and Location Optimization", *IEEE Communication Letters*, vol. 25, no., 1, pp. 269--273, DOI:10.1109/LCOMM.2020.3025345, 2021.

Singh, R. and Sicker, D., "THz Communications - a Boon and/or Bane for Security, Privacy, and National Security", *TPRC48: The 48th Research Conference on Communication, Information and INternet Policy,* 2020.

Soderi, S., and De Nicola, R., "6G Networks Physical Layer Security Using RGB Visible Light Communications", *IEEE Access*, vol. 10, pp. 5482--5496, DOI: 10.1109/ACCESS.2021.3139456, 2021.

Ziegler, V., Schneider, P., Viswanathan, H., Montag, M., Kanugovi, S. and Rezaki, A., "Security and Trust in the 6G Era", *IEEE Access*, vol. 9, pp.142314--142327, DOI: 10.1109/ACCESS.2021.3120143, 2021.

Tackling Uncertainty Through Probabilistic Modelling of Proportionality in Military Operations

Clara Maathuis[1] and Sabarathinam Chockalingam[2]
[1]Open University of the Netherlands, Heerlen, Netherlands
[2]Institute for Energy Technology, Halden, Norway

clara.maathuis@ou.nl
sabarathinam.chockalingam@ife.no

Abstract: Just as every neuron in a biological neural network is a reinforcement learning agent, thus a component of a large and advanced structure is de facto a model, the two main components forming the principle of proportionality in military operations can be seen and are as a matter of fact two different entities and models. These are collateral damage depicting the unintentional effects affecting civilians and civilian objects, and military advantage symbolizing the intentional effects contributing to achieving the military objectives defined for military operation conducted. These two entities are complex processes relying on available information, projection on time to the moment of target engagement through estimation and are strongly dependent of common-sense reasoning and decision making. As a deduction, these two components and the proportionality decision result are processes surrounded by various sources and types of uncertainty. However, the existing academic and practitioner efforts in understanding the meaning, dimensions, and implications of the proportionality principle are considering military-legal and ethical lenses, and less technical ones. Accordingly, this research calls for a movement from the existing vision of interpreting proportionality in a possibilistic way to a probabilistic way. Henceforth, this research aims to build two probabilistic Machine Learning models based on Bayesian Belief Networks for assessing proportionality in military operations. The first model embeds a binary classification approach assessing if the engagement is proportional or disproportional, and the second model that extends this perspective based on previous research to perform multi-class classification for assessing degrees of proportionality. To accomplish this objective, this research follows the Design Science Research methodology and conducts an extensive literature for building and demonstrating the model proposed. Finally, this research intends to contribute to designing and developing explainable and responsible intelligent solutions that support human-based military targeting decision-making processes involved when building and conducting military operations.

Keywords: Machine Learning, Bayesian Networks, Military Operations, Cyber Operations, Targeting, Proportionality.

1. Introduction

"*Peace cannot be kept by force. It can only be achieved by understanding*." (Albert Einstein)

Ongoing wars demonstrate that although significant technological developments have been made in all societal domains as well as in the military domain, still preventing, limiting, or containing the unintended effects on civilian and civilian objects while achieving military goals, is difficult. This happens due to various sources and types of uncertainty surrounding entities like the context, enemy, target and corresponding intelligence information, and environment plus the multiple dimensions characterizing (in one way or another) the humans involved, e.g., brittleness, bias, unintelligibility, and subjectivity (Margulies, 2020; Morgan et al., 2020). At the core of this phenomenon (i.e., war) is the military targeting process which implies selecting, prioritizing, and matching targets with means and methods for achieving goals and projecting force on the enemy (US Army, 2013; NATO, 2016). A central feature characterizing this process is the proportionality principle directly expressed through the proportionality assessment when conducting military operations (Crow, 2019). In this assessment, two critical and antagonist components (Henderson & Reece, 2018) representing the two parts behind the wall: the anticipated military advantage as intended effects that contribute to achieving military objectives and the estimated collateral damage as unintended effects on civilians and civilian objects (Maathuis, Pieters & van den Berg, 2018b), are brought together so that the military Commander can determine and make sure that collateral damage would not be excessive in relation to military advantage. The proportionality assessment process is in fact a collective decision-making process where multiple agents are involved and advise the military Commander through their input and expertise, for establishing and making proper and legal targeting decisions (Holland, 2002; Colonomos, 2017; Bartneck et al., 2021).

A large body of academic and practitioner studies exist with respect to understanding and applying different perspectives on important concepts that represent a key element when planning, conducting, and assessing military operations, e.g., attack, military objective, target identification, and proportionality. Whereas research on capturing and representing such knowledge for modelling and simulating them using intelligent techniques is in its infancy. To this end, this research aims to capture, represent, and model the components involved when

conducting the proportionality assessment in military operations by designing and developing two probabilistic Machine Learning models using Bayesian Belief Networks. On this, Morgan et al., (2020) and Boury-Brisset & Berger (2020) emphasize the benefits of building and using Artificial Intelligence / Machine Learning models for military decision-making support. Accordingly, to the best of our knowledge, these models represent the first attempt in this sense in the existing body of knowledge to move from the possibilistic approach to the probabilistic approach and aim to directly support military targeting decision-making processes. On these grounds, a Design Science Research methodological approach (Peffers, Tuunanen & Niehaves, 2018) is considered and followed to reach the stated goal while having the following contributions:

- Awareness and support to military decision makers such as military Commanders, military-technical, and military-legal experts involved when planning and executing military operations when dealing with different uncertainty dimensions, sources, and entities characterizing the proportionality assessment process.
- A direct call for future research in AI-based and gaming-based modelling and simulation for military targeting decision-making support based on multidisciplinary perspectives, methods, techniques, and technologies for building efficient, effective, trustable, and accountable military operations.

The remainder of this article is structured as follows. In Section 2, the context of this research is discussed together with related studies. In Section 3, the research methodology pursuit to achieve the goal of this research is tackled. In Section 4, the design of the two models proposed is presented together with important design components and choices. In Section 5, the evaluation mechanism considered is elaborated through demonstration on two military operations scenarios and comparison with another relevant state-of-the-art AI model in this field. At the end, in Section 6, reflection, concluding remarks, and future ideas are discussed.

2. Background and Related Research

To postulate the background of this research, it is important to mention the existence of three theoretical visions in relation to war: *jus ad bellum* that refers to the principles governing the justice of going to war, *jus in bello* that refers to the ones governing the conduct of warfare, and just *post bellum* that refers to the ones after war (Stahn, 2006; Pavlischek, 2010). Given the fact that this research positions itself in time of war it means that it finds itself under the *jus in bello* principles of conducting war, specifically referring to targeting in both international and non-international conflicts (Boothby & Schmitt, 2012; Tallinn Manual, 2017). The core of this setting is the conduct of military operations for influencing targets in different ways, e.g., communication disruption or alter the behaviour of an audience, captured in the military targeting process which implies the selection, prioritization, and match of targets to corresponding responses while considering operational requirements and capabilities, e.g., ROEs (Rules of Engagement) (NATO, 2016; US Army, 2013). To conduct such operations, agents (i.e., state/non-state actors) use means and methods of warfare that "are indispensable" (Downey, 1953) to obtain military advantage against their adversaries and achieve their aim (principle of military necessity). For this, the agents need to make a clear distinction between "those who may be lawfully attacked and those who must be respected and protected" (Boothby & Schmitt, 2012), in other words, between combatants and military objects, and civilian objects and humans, respectively (principle of distinction). For this, multiple aspects are established, and decisions are taken considering (i) if the intended target is a human or an object, and (ii) in respect to different criteria, deciding if the human or object is a lawful target that contributes to the achievement of military aims (Whittemore, 2015). Hence, military necessity shows the attempt to realize the aim through the gain of military advantage that must make an effective contribution to the military action and be concrete and perceptible (Harutyunyan, 2019). At the same time humanitarian considerations aim to minimize human suffering and physical destruction (Wallace & Reeves, 2019). This represents the basis for assessing that an attack that can "cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated" is disproportional, thus forbidden (AP I Art.51(5)(b), 1977, principle of proportionality). In its essence, the principle of proportionality brings in balance two antagonistic concepts (Maathuis, C., Pieters, W., & van den Berg, 2018a), based on "timely, accurate, and reliable information" available at that time (US Army, 2013). Nevertheless, this process is surrounded by different sources of uncertainty regarding aspects such as (i) relative values to be assigned to the two concepts, (ii) what can be included and excluded when summing the two concepts, (iii) which time and space measurement standard should be used, and (iv) how much should be the military forces conducting the operation exposed for limiting civilian casualties or damage to civilian objects (Holland, 2002). Shortly, Colonomos (2017) refers to it as being "a compromise between political constraints, military necessities, legal claims, and ethical aspirations".

Crow (2019) discusses three distinct approaches for understanding proportionality: as a legal principle, as a goal of government, and a particular structured approach to judicial review. Accordingly, the first and third approaches combined result into a legal obligation conducted in a structured approach characterized by aspects such as legitimacy or suitability, reasonableness or rationality, necessity, and balancing. Particularly, reasonableness represents "an objective standard for reasonability" which implies that a set of constructed values and norms are projected and applied during the assessment. Moreover, Henderson & Reece (2018) analyze what the 'reasonable military Commander' means and implies in the context of proportionality assessment in military operations. In this sense, the assessment is compared to a general assessment that can be seen as subjective (the person believes), objective but unqualified (the person reasonably believes), or objective but qualified (the doctor, i.e., expert, reasonably believes). For realizing a proper assessment, they stress the importance of military training and active participation in military exercises.

Gillard (2018) emphasizes the importance of understanding the elements involved as well as the underlying aspects contributing to the proportionality assessment. On this behalf, the author stresses firstly the relation between *causation* and *foreseeability*, i.e., the expected effects caused by the attack and the fact that the assessment itself is *ex ante*, i.e., conducted before the attack takes place. In addition to these terms, Gillard (2018) refers to "reasonable foreseeability, i.e., what should have been foreseen by the party responsible at the time of the act" implying that "what can reasonably be foreseen depends on the circumstances in which the attack is planned, decided, or launched". Secondly, the author emphasizes and relates the expectancy of effects to likelihood which means the probability that such effects will occur. And thirdly, the author discusses the weighting perspective, i.e., "once the incidental harm to be considered in a proportionality assessment has been identified, a value or weight must be assigned to it" and the *excessive* element. Mathematically, this can be directly translated to a weight parameter assigned with different values to the concepts involved in this equation to describe the proportionality function or as a weight parameter assigned again with different values while searching to balance the participating components in this equation.

In addition to the uncertainty sources identified above, Katzir (2018) and Morgan et al. (2020) add the subjectivity of the proportionality assessment process. At the same time, Morgan et al. (2020) considers the proportionality assessment to be "an evaluative, qualitative, and ethical assessment [done] by a human weighting and comparing complex values". Moreover, other sources of uncertainty can be added based on heuristics and biases surrounding this process (Whittemore, 2015) born from facts like limitations imposed to cognitive capacities (e.g., time, space, and cost), complexity of the environment, and decision-making impediments due to perceived irrationality. These sources could not only increase the complexity and duration of the assessment but could also lead to an improper decision. In case of disproportionality, military Commanders are tried either because civilians instead of combatants were targeted or because the force applied led to an excessive collateral damage (Maathuis, C., Pieters & van den Berg, J. (2021).

As the background and related studies show, significant attention is provided on understanding the elements involved in the principle of proportionality in the body of knowledge while a limited number of studies is dedicated to designing and developing solutions to model it by capturing various sources and/or types of uncertainty. To this end, it is the aim of this article to capture and represent uncertainty in a probabilistic way building from scratch two BBN-based Machine Learning models as described in the following sections of this article.

3. Research Methodology

This research aims to capture and represent uncertainty present when conducting the proportionality assessment in military operations by means of building two Bayesian Belief Network models for providing decision support when targeting in military operations. Accordingly, this research intends to answer the following research questions:

- How to represent uncertainty surrounding the components involved in the proportionality assessment in military operations?
- How to build probabilistic models that captures uncertainty surrounding the components involved in the proportionality assessment in military operations?

To this end, a multidisciplinary stance is considered taking a Design Science Research methodological approach in an explanatory, predictive, and probabilistic way pursuing the following research activities (Kuechler & Vaishnavi, 2012; Peffers, Tuunanen & Niehaves, 2018):

In Phase I, *Problem statement and aim definition*, the background, disciplines, and dimensions involved and characterizing this research are established. Furthermore, taking into consideration the fact that in general a subjective possibilistic approach is considered when assessing proportionality in military operations and the fact that as the literature and field experts suggest, that it is a process surrounded by uncertainty, this research considers a probabilistic approach for the proportionality assessment building on previous work by (Maathuis, Pieters & van den Berg, 2018b; Maathuis, Pieters & van den Berg, 2021; Maathuis & Chockalingam, 2023). This represents to the best of our knowledge the first research effort in this direction. The review of relevant research studies is conducted using different combinations of keywords like 'proportionality', 'test', 'assessment', 'military targeting', and 'military operations' for querying scientific databases such as IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar.

In Phase II, *Solution development*, the design of the models proposed is established as follows: the first model captures a direct probabilistic approach considering that the engagement of a specific target can be either proportional or disproportional together with associated probabilities, and the first model captures an extended probabilistic approach by introducing different levels of proportionality when engaging a specific target, on other words, from very low disproportional to proportional. Once the variables, states, and rules of the models are designed, they are further implemented as Bayesian Belief Networks.

In Phase III, *Solution evaluation and Communication*, the models proposed are evaluated considering synthetical military operations scenarios published in dedicated venues in this domain. Specifically, the evaluation is conducted through exemplification, i.e., both models proposed are exemplified on two military operations scenarios. Furthermore, the research conducted, models proposed, and results obtained are communicated through this article as well as corresponding presentations in different scientific settings.

4. Models Design

The process of preparing and executing the proportionality assessment can be reduced to identifying the target, outlining the anticipated military advantage, and expected collateral damage, further indicating that the military Commander has determined that under the law engaging the target would be proportional or disproportional, and further requesting final advice in case that other aspects should be considered as well (Henderson & Reece, 2018). Accordingly, two perspectives are considered when modelling the principle of proportionality considering its two main components:

In the first perspective (direct) that corresponds to the first model, the two input variables military advantage (MA) and collateral damage (CD) contain three states: Low (L), Medium (M), and High (H), while the target variable is the proportionality assessment (PA) which has two states: Proportional (P) and Disproportional (DP). To exemplify, the logic behind the first model is captured in equation (1):

If CD is Low AND MA is High, THEN PA is Proportional
If CD is High AND MA is Medium, THEN PA is Disproportional *(1)*

In the second perspective (extended) that corresponds to the second model, the two input variables military advantage (MA) and collateral damage (CD) contain five states: Very Low (VL), Low (L), Medium (M), High (H), and Very High (VH), while the target variable is the proportionality assessment (PA) which has four states: Very High Disproportional (VHDP), High Disproportional (HDP), Low Disproportional (LDP), and Proportional (P). As exemplification, the logic behind the second model is presented in equation (2):

If CD is Very Low AND MA is High, THEN PA is Proportional
If CD is High AND MA is Low, THEN PA is Very High Disproportional *(2)*

Furthermore, the design of both models are implemented using Bayesian Belief Networks seeing their capacity to model complex systems, processes, and activities under uncertainty in different societal domains (Kahn Jr et al., 1997; Marcot & Penman, 2019). Accordingly, both models contain two main components. The first component is qualitative and is represented by a Directed Acyclic Graph (DAG) as shown in Figure 1 and Figure 2. that embodies the variables (input and target) and their corresponding states. And the second component is quantitative, which is represented by the Conditional Probabilities (CPTs) of variables containing the conditional probabilities of all combinations that are possible for child-parent variable states, as depicted in Figure 1 and Figure 2. In case that a variable does not have a parent variable, its CPT contains the priori marginal probabilities of the corresponding variable. Moreover, the reasoning considered in this research is predictive being done from cause (upper layer) to effect (lower layer) (Maathuis & Chockalingam, 2023) and the values illustrated in Figure 1 and Figure 2 are without evidence provided, i.e., for showing the use of the models which are further evaluated

through demonstration in Section 5.

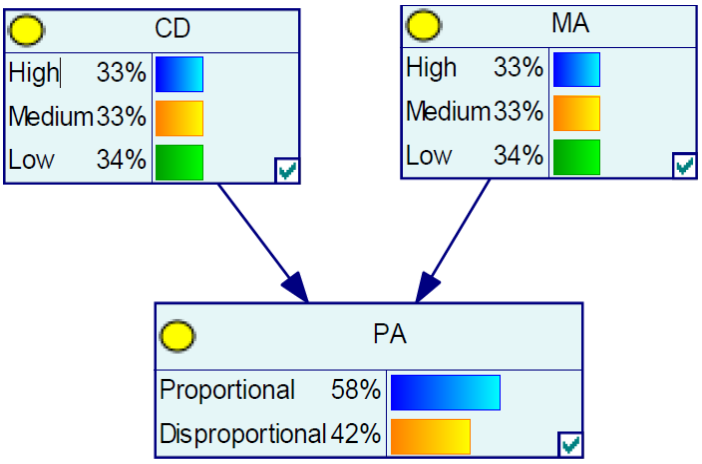


Figure 1. First model

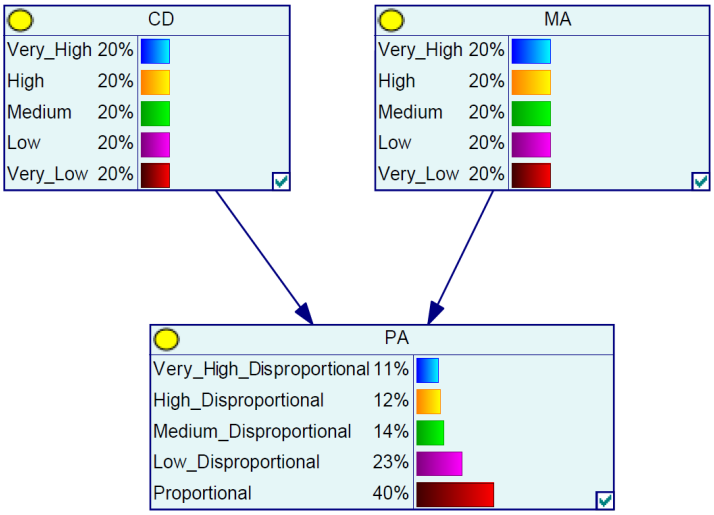


Figure 2. Second model

5. Model Evaluation

The evaluation of the models proposed is done through demonstration based on two military operations scenarios that contain synthetical data used for evaluating a deep multi-layered fuzzy model proposed for proportionality assessment in military Cyber Operations by Maathuis, Pieters & van den Berg (2021). Accordingly, the results of the models proposed in this research are compared with the ones obtained by the multi-layered fuzzy model built in previous work. This comparison is done in order to evaluate the reasoning of the models proposed in this article. Moreover, a short description of the two scenarios is provided together with the demonstration of the models proposed in this research. Hence, the models proposed in this research receive as input for the variables contained in the data from the multi-layered model and the proportionality assessment decision is then compared next to the decision provided by consulted military experts.

In the first scenario, a drone military Cyber Operation is planned to prevent a terrorist attack using a suicide drone/unmanned combat aerial vehicle (UCAV) weaponized with 3 kg explosive munition against the president of a country. For this, the position and speed of the drone is altered in the ground control station. Hence, the proportionality assessment should be conducted regarding executing the corresponding cyber weapon that will do this action. In this case, the multi-layered model estimated a high collateral damage and a medium military advantage. Then, these values are provided as input for the models proposed in this research as depicted in Figure 3 and Figure 4, and further compared in Table 1.

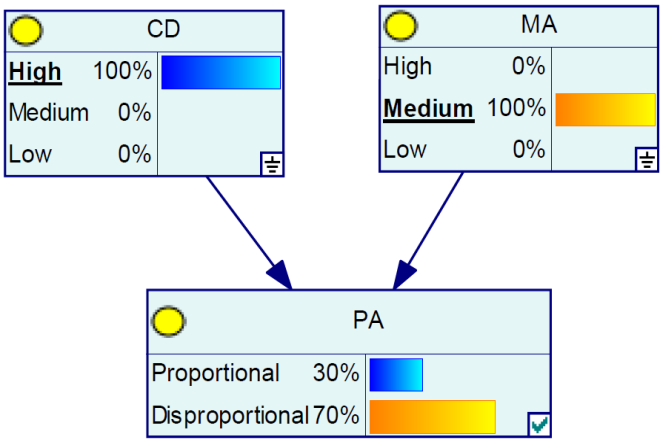


Figure 3. First model demonstration with the first scenario

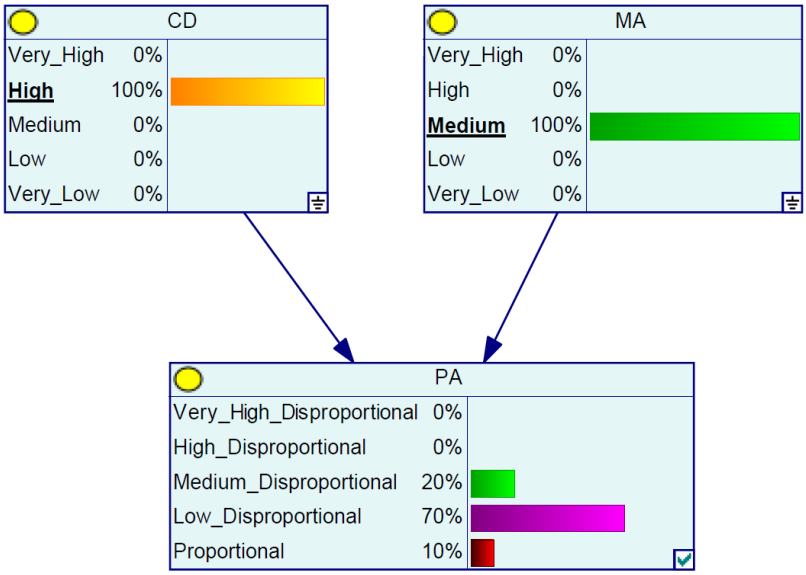


Figure 4. Second model demonstration with the first scenario

In the second scenario, a ship military Cyber Operation is planned to prevent a terrorist attack using a commercial cargo ship weaponized with chemical agents. For this, a protocol-based DDoS is prepared against the pump station by exploiting an unpatched software vulnerability. In this case, the multi-layered model estimated a medium collateral damage and a medium military. Then, these values are provided as input for the models proposed as depicted in Figure 5 and Figure 6.

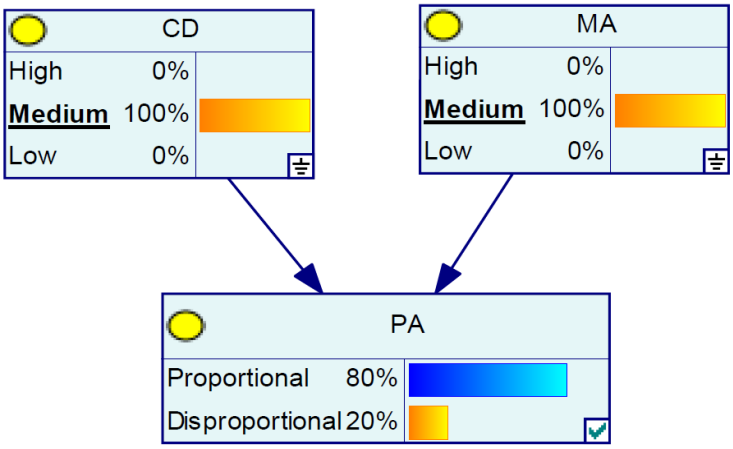


Figure 5. First model demonstration with the second scenario

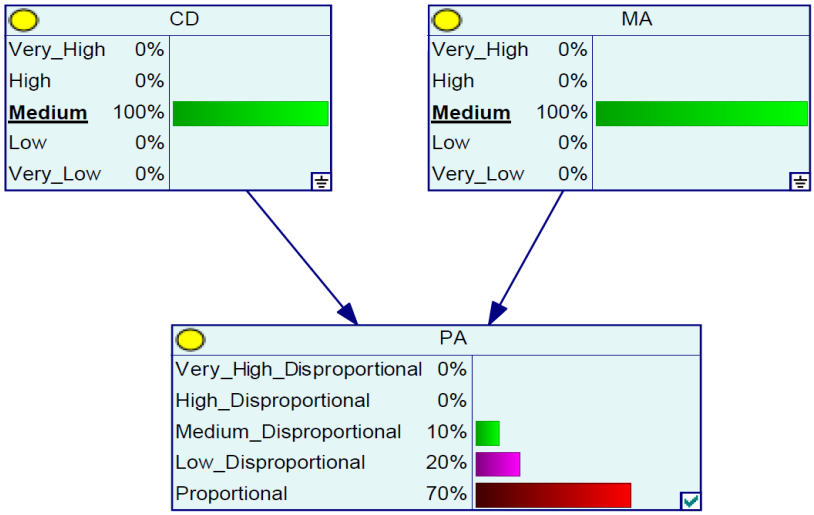


Figure 6. Second model demonstration with the second scenario

Henceforth, the values obtained by the model proposed in this research are compared in Table 1 with the ones obtained by the multi-layered model proposed and evaluated with military experts by Maathuis, Pieters & van den Berg (2021).

Conclusively, the results show that the models proposed in this research provide a probabilistic advancement for the existing model in this domain. Accordingly, the highest probabilities assessed by both models herein match the decision provided by the existing AI model in both cases of military operations examples. Additionally, these assessments match the decisions provided by experts consulted in this field in previous research. Hence, the models proposed are evaluated through comparison and the results of this evaluation show that they represent a realistic solution for capturing and modelling uncertainty dimensions when assessing proportionality in military operations.

Table 1. Comparison of models and assessment of experts

Military Operation Scenario	Proportionality Assessment Expert Decision 1, 2, 3, 4	Multi-layered model Proportionality Assessment	First model Proportionality Assessment	Second model Proportionality Assessment
1	Proportional Disproportional Disproportional Disproportional	Disproportional	30 % Proportional 70 % Disproportional	0 % Very High Disproportional 0 % High Disproportional 20% Medium Disproportional 70 % Low Disproportional 10 % Proportional
2	Proportional Proportional Proportional Proportional	Proportional	80 % Proportional 20 % Disproportional	0 % Very High Disproportional 0 % High Disproportional 10% Medium Disproportional 20 % Low Disproportional 70 % Proportional

6. Conclusions

Wars are a phenomenon characterizing the human existence and experience since the dawn of history (Tabansky, 2011). As they directly represent the planning, execution, and assessment of military operations for reaching (pre)defined goals, important legal, ethical, and social aspects need to be considered for minimizing harm to civilians and protecting combatants based on previous experience. Additionally, the technological development is beneficial (Pfaff, 2016) since it allows building the capacity of (effectively and efficiently) modelling and simulating different activities, actions, and aspects. Among these can be mentioned the achievement of objectives established and the effects produced while providing accurate predictions having different degrees of autonomy for military decision-making support. In this realm, Artificial Intelligence/Machine

Learning techniques and applications such as image recognition, natural language processing, and voice recognition already showed impressive results in target identification and engagement control (Panwar, 2017; Bartneck et al., 2021).

At the core of this phenomenon is the principle of proportionality. Nevertheless, while a vast number of studies concerning the meaning and applicability of the principle of proportionality in different military operations domains and scenarios are tackled in the existing body of knowledge, a limited number of studies is dedicated to capturing, representing, and modelling the proportionality assessment in military operations. To this end, this research aims to model the proportionality assessment in military operations by capturing and representing its components translated to variables and values taking a probabilistic approach using ML. This is done by building two Bayesian Belief Network models. The first model contains a direct perspective on the proportionality decision while the second model takes an extended perspective on it based on the need expressed both in the literature and by military-technical experts for considering different levels of proportionality. This approach is adopted as it is a well-known approach for tackling uncertainty and to reason in complex, uncertain, ill-defined, and subjective domains, and tasks such as the proportionality assessment. To do this, a multidisciplinary research is conducted merging military, legal, and AI knowledge and expertise following a Design Science Research approach (Peffers, Tuunanen & Niehaves, 2018). Accordingly, the models proposed are further evaluated through demonstration on two military operations scenarios and compared with a previous AI model and proportionality assessment conducted by military experts. The results obtained show the fact that such an approach represents a proper way of capturing, representing, and modelling uncertainty dimensions surrounding different components/dimensions of the principle of proportionality in military operations.

As this research represents the first attempt in the existing body of knowledge to consider a direct probabilistic approach to the principle of proportionality in military operations, it aims to further integrate multiple classes of variables that characterize each of the two components. Further, it can be extended with multiple data sources and multiple clusters of effects while evaluating it using different types of military operations carried out in different domains using both real and synthetical data. In this way, this research paves the way to build explainable, responsible, and trustable intelligent solutions (Chockalingam & Maathuis, 2022; Maathuis, 2022a; Maathuis, 2022b) for supporting planning, execution, and assessment of (digitally prepared) military operations.

References

Additional Protocol I (1977). Art 51(5)(b) – Protection of the civilian population.

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). Military Uses of AI. In *An Introduction to Ethics in Robotics and AI* (pp. 93-99). Springer.

Boury-Brisset, A. C., & Berger, J. (2020). Benefits and Challenges of AI/ML in Support of Intelligence and Targeting in Hybrid Military Operations. NATO S&I Organisation.

Boothby, W., H. & Schmitt, M., N. (2012). *The law of targeting*. Oxford University Press. Pavlischek, K. (2010). Proportionality in warfare. *The New Atlantis*, (27), 21-34.

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).

Colonomos, A. (2017). Proportionality in Warfare as a Political Norm. *Weighing Lives in Warfare*, 217-240.)

Crow, K. (2019). The Opacity of Proportionality in International Courts: Could Categories Clarify?. *Geo. Wash. Int'l L. Rev.*, *51*, 289.

Downey, W. G. (1953). "The law of war and military necessity." *American Journal of International Law* 47.2: 251-262.

Gillard, E. C. (2018). Proportionality in the conduct of hostilities: the incidental harm side of the assessment.

Henderson, I., & Reece, K. (2018). Proportionality under International Humanitarian Law: The Reasonable Military Commander Standard and Reverberating Effects. *Vand. J. Transnat'l L.*, *51*, 835.

Harutyunyan, A. (2019). Dilemma of Targeting: Dual-Use Objects in Military Operations. *Law of Armed Conflict*.

Holland, J. C. (2002). *Military Objective and Collateral Damage: Their Dynamics and Relationship*. JUDGE ADVOCATE GENERAL'S SCHOOL CHARLOTTESVILLE VA.

Kahn Jr, C. E., Roberts, L. M., Shaffer, K. A., & Haddawy, P. (1997). Construction of a Bayesian network for mammographic diagnosis of breast cancer. Computers in biology and medicine, 27(1), 19-29.

Henderson, I., & Reece, K. (2018). Proportionality under International Humanitarian Law: The Reasonable Military Commander Standard and Reverberating Effects. *Vand. J. Transnat'l L.*, *51*, 835.

Holland, J. C. (2002). *Military Objective and Collateral Damage: Their Dynamics and Relationship*. JUDGE ADVOCATE GENERAL'S SCHOOL CHARLOTTESVILLE VA.

Katzir, R. (2018). Four Comments on the Application of Proportionality under the Law of Armed Conflict. *Vand. J. Transnat'l L.*, *51*, 857.

Kuechler, W., & Vaishnavi, V. (2012). A framework for theory development in design science research: multiple perspectives. *Journal of the Association for Information systems*, *13*(6), 3.

Maathuis, C., Pieters, W., & van den Berg, J. (2018a). A knowledge-based model for assessing the effects of cyber warfare. In *Proceedings of the 12th NATO Conference on Operations Research and Analysis*.

Maathuis, C., Pieters, W., & van den Berg, J. (2018b). Developing a cyber operations computational ontology. *Journal of Information Warfare*, *17*(3), 32-49.

Maathuis, C., Pieters, W., & van den Berg, J. (2021). Decision support model for effects estimation and proportionality assessment for targeting in cyber operations. *Defence Technology*, *17*(2), 352-374.

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. & Chockalingam, S. (2023). Modelling the influential factors embedded in the proportionality assessment in military operations. In *International Conference on Cyber Warfare and Security*.

Marcot, B. G., & Penman, T. D. (2019). Advances in Bayesian network modelling: Integration of modelling technologies. *Environmental modelling & software*, *111*, 386-393.

Margulies, P. (2020). Autonomous Cyber Capabilities Below and Above the Use of Force Threshold: Balancing Proportionality and the Need for Speed. *International Law Studies*, *96*(1), 13.

Morgan, F. E., Boudreaux, B., Lohn, A. J., Ashby, M., Curriden, C., Klima, K., & Grossman, D. (2020). *Military applications of artificial intelligence: ethical concerns in an uncertain world*. RAND PROJECT AIR FORCE SANTA MONICA CA SANTA MONICA United States.

NATO (2016). *NATO Standard AJP-3.9 Allied Joint Doctrine for Joint Targeting*. NATO Standardization Office.

Panwar (2017). Artificial Intelligence in Military Operations: an overview – Part I. https://futurewars.rspanwar.net/artificial-intelligence-in-military-operations-an-overview-part-i/

Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research.

Pfaff, C. A. (2016). Five myths about military ethics. *The US Army War College Quarterly: Parameters*, *46*(3), 8.

Schmitt, M. N. (Ed.). (2017). *Tallinn manual 2.0 on the international law applicable to cyber operations*. Cambridge University Press.

Stahn, C. (2006). 'Jus ad bellum','jus in bello'...'jus post bellum'?–Rethinking the Conception of the Law of Armed Force. *The European Journal of International Law*, *17*(5), 921-943.

Tabansky, L. (2011). Basic concepts in cyber warfare. *Military and Strategic Affairs*, *3*(1), 75-92.

United States Army (2013). *Joint Publication 3-60 Joint Targeting*. United States Army.

Wallace, D. A., & Reeves, S. R. (2019). Protecting Critical Infrastructure in Cyber Warfare: Is It Time for States to Reassert Themselves?. *UC Davis L. Rev.*, *53*, 1607.

Whittemore, L. A. (2015). Proportionality Decision Making in Targeting: Heuristics, Cognitive Biases, and the Law. *Harvard National Security Journal*, *7*, 577.

Whittemore, L. A. (2015). Proportionality Decision Making in Targeting: Heuristics, Cognitive Biases, and the Law. *Harv. Nat'l Sec. J.*, *7*, 577.

Design Lessons from Building Deep Learning Disinformation Generation and Detection Solutions

Clara Maathuis[1], Iddo Kerkhof[2], Rik Godschalk[3] and Harrie Passier[4]
[1,4] Open University, Netherlands
[2,3] Independent Researcher, Netherlands

clara.maathuis@ou.nl
email@iddo.eu
ahgodschalk@gmail.com
harrie.passier@ou.nl

Abstract: In its essence, social media is on its way of representing the superposition of all digital representations of human concepts, ideas, believes, attitudes, and experiences. In this realm, the information is not only shared, but also {mis, dis}interpreted either unintentionally or intentionally guided by (some kind of) awareness, uncertainty, or offensive purposes. This can produce implications and consequences such as societal and political polarization, and influence or alter human behaviour and beliefs. To tackle these issues corresponding to social media manipulation mechanisms like disinformation and misinformation, a diverse palette of efforts represented by governmental and social media platforms strategies, policies, and methods plus academic and independent studies and solutions are proposed. However, such solutions are based on a technical standpoint mainly on gaming or AI-based techniques and technologies, but often only consider the defender's perspective and address in a limited way the social perspective of this phenomenon becoming single angled. To address these issues, this research combines the defenders' perspective with the one of the offenders by (i) building a hybrid deep learning disinformation generation and detection model and (ii) capturing and proposing a set of design recommendations that could be considered when establishing patterns, requirements, and features for building future gaming and AI-based solutions for combating social media manipulation mechanisms. This is done using the Design Science Research methodology in Data Science approach aiming at enhancing security awareness and resilience against social media manipulation.

Keywords: social media manipulation, disinformation, misinformation, security awareness, machine learning, deep learning.

1. Introduction

"*We live in a fantasy world, a world of illusion. The great task in life is to find reality*." (Iris Murdoch)

Through their capability of producing confusion, deception, and polarization by influencing human beliefs and behaviour, social media manipulation mechanisms like disinformation and misinformation were recently positioned among the most impactful cyber threats (Caramancion et al., 2022; ENISA, 2022). Such mechanisms have attested records from millennia's ago in the disinformation campaigns conducted in Egypt surrounding the Battle of Kadesh where a wrong location was given to Pharaoh Ramses II, to more recent collection of articles published by The Sun named the Great Moon Hoax in 1835 about the life and culture found on the Moon, and during the World Wars regarding the demonization of parties involved and propaganda (Choras et al., 2021). Due to technological advancements, such mechanisms evolved rapidly reaching global audiences while their counterparts also advance in building combating solutions based on content/data, (creative and participatory) engagement, and a broad range of software and AI-based solutions (Reisach, 2021; Maathuis, 2022b). Nevertheless, social media manipulation is an online phenomenon with offline impact (Shu, 2022). Corresponding combat efforts usually take the defenders' perspective, but a paradigm switch and change of approach to considering attackers' perspective is necessary, but scarcely present in the literature and practitioner efforts.

Hence, this research aims to build a hybrid deep-learning model for generating and detecting disinformation based on previous work, and from there to propose a set of design recommendations that can serve as design patterns, requirements, and features for building responsible software (e.g., gaming) and AI-based solutions for producing or enhancing awareness, resilience, and education on social media manipulation. To achieve this aim, a multidisciplinary stance is taken following the Design Science Research in a Data Science methodological approach (Peffers et al., 2007; Peffers, Tuunanen & Niehaves, 2018; Foroughi & Luksch, 2018) having the following contributions:

- From an implementation perspective, it proposes a hybrid intelligent solution that merges the perspective of the attackers with the defenders' one using advanced deep learning techniques emphasizing the possibility of combining real with synthetic content and obtaining relevant results.

- From a design perspective, it proposes a set of recommendations with socio-technical dimensions that could be integrated when building software-based or intelligent solutions for combating social media manipulation.

This article advances as follows. Section 2 settles the background and presents relevant studies to this research. Section 3 addresses the research methodology pursued to achieve the aim of this research. Section 4 summarizes the deep learning techniques used to build the hybrid model proposed. Section 5 presents important implementation aspects and the evaluation results obtained. Section 6 addresses the recommendations proposed. At the end, Section 7 presents concluding remarks and future ideas.

2. Background and Related Research

In its last report, ENISA (European Union Agency for Cybersecurity) positions social media manipulation among the top eight cyber threats of 2021 and 2022 (ENISA, 2022). To deal with it, Reisach (2021) considers that all agents involved, i.e., social media platforms, organizations, users, society, share responsibility. When aiming at limiting/combating social media manipulation, it is necessary to realize the importance of awareness and education. Hence, corresponding solutions that build a resilient and self-aware social media literacy and mindset should be proposed. Moreover, Interpol (2020) argues for building accurate and trustworthy AI solutions for tackling mis/disinformation mechanisms while accounting existing policy, legal, and societal challenges (Europol, 2022). As a basis for building such solutions whilst protecting human and democratic values and systems, limiting public harm and damage on policy-making processes while contributing to building users' digital security and resilience, the EU Commission proposed initiatives like European Democracy Action Plan and the Strengthened Code of Practice on Disinformation (EU Commission, 2018; EU Commission, 2022a; EU Commission, 2022b).

Thakur, Hayajneh & Tseng (2019) analyze cyber threats in social media together with possible ways to counter them and limit the damage. Among these threats are social engineering attacks and lack of social media policies. Social engineering attacks are based on eliciting individuals' personal information using fake accounts and building trust. Lack of social media policies for organizations points to guidelines that should be deployed for safe social media use by its employees based on proper education and training plus the establishment of a social media coordinator, champion, or manager. Shu et al. (2020) discuss differences and connections between different manipulation mechanisms and emphasize the importance of user engagement, proper design of detecting and mitigating solutions, and the inclusion of socio-technical aspects and technological paradigms like ethics and blockchain, respectively. Moreover, Zhou & Zafarani (2020) conduct a systematic literature review study on fake news theories, detection methods, and opportunities for detection purposes. The authors categorize the studies as follows: the fake knowledge carried, writing style, propagation patterns, and credibility of the source. On this behalf, the importance of {multi, inter, trans}disciplinary research and collaboration between experts from different disciplines which would conduct to explainable results, is highlighted. Choaras et al. (2021) review advanced Machine Learning applications using existing datasets concentrating on scientific and R&D societal projects. The results show the importance of understanding the context, having a clear methodological approach, and being able to explain both the modelling reasoning and the results obtained. Additionally, Pinhanez et al. (2022) argue that disinformation, and particularly, fake news and deep fakes, are an important upcoming cyber security problem that could be tackled building a new way of understanding disinformation by merging generation and detection perspectives while actively engaging users in prevention and treatment processes. On this, Rogers & Niederer (2020) analyze political dimensions of politically deployed social media manipulation and identifies two major waves when dealing with this. The first regarding understanding, i.e., definitional issues and production, and the second dedicated to its consumption which includes persuasion. In both waves, AI-based efforts for detecting and dealing with such mechanisms exist. Alassad, Spann & Agarwal (2021) propose an integrated method for identifying malicious behaviour and actors responsible for propagation based on advanced computational social science methods and graph theoretic techniques. The authors point the importance of user versus group analysis for detecting and investigating malicious intent and provide further decision-making support for applying proper response strategies. Moreover, Reisach (2021) advances a responsibility-based approach for social media platforms with a keen eye on political contexts based on existing socio-ethical principles and requirements developed for safeguarding democratic societies through human rights consideration, education, and transparency of algorithmic decisions. Kertysova (2018) investigates challenges and opportunities posed by AI in the context of producing, disseminating, and countering disinformation. Particularly, the author stresses the importance of understanding and dealing with both human-based and algorithmic-based biases when dealing with disinformation campaigns/operations, e.g., cultural, racial, gender, personality, and political, which improperly treated could

imply disparate impact on certain groups of people. In respect to these dimensions, Maathuis & Chockalingam (2022a) propose a model that incorporates human dimensions for assessing if the digital security behaviour of users could be classified as responsible or not.

Furhermore, Nasir, Khan & Varlamis (2021) propose a hybrid deep learning based model combining CNN (Convolutional Neural Network) with RNN (Recurrent Neural Network) techniques for fake news detection showing promising generalizing results. Stiff & Johansson (2022) propose a RoBERTa-based disinformation detection solution which is robust on both in-distribution and out-of-distribution test data plus realistic in-the-wild data. Kaliyar, Goswami & Narang (2021) build a hybrid deep learning fake news detection model combining BERT (Bidirectional Encoder Representations from Transformers) with CNN techniques for capturing information relevance and reducing ambiguity. Given the approach pursued and results obtained (accuracy of 98.99%), in this research a similar approach is considered for the detection component of the solution proposed. Verma et al. (2022) build the MCred (Message Credibility) framework based on CNN and BERT techniques for disinformation detection through information credibility assessment using the benefits of local and global text semantics. Du, Bosselut & Manning (2022) build a deep learning-based verification framework for sensitivity evaluation of deep learning-based techniques, i.e., KGAT, CorefBERT, and MLA to adversarial generated disinformation done using GROVER. The results obtained show a high sensitivity to automated generated disinformation on different topics. This implies that improvements in automated synthetic content detection is necessary and should be combined with human audits of fact-checker evidence, thus a joint between algorithm-based modelling and expert-based approach. Shu, Li, Ding & Liu (2021) propose a solution for news generation based on FACTGEN (Fact-Enhanced Synthetic News Generation) method that is implementing a PSA (Pseudo-Self-Attentive) model for tackling challenges related to information encountered when generating news on a given claim. Furthermore, it is important to acknowledge the importance of security and privacy threats to deep learning models applied in this domain. Le, Wang & Lee (2020) build a GAN (Generative Adversarial Network) solution for generating malicious content for attacking fake news detection models. Bradshaw & Howard (2018) conduct an extensive literature review, consult field experts, and use content analysis solutions on social media manipulation incidents conducted in 48 countries on Facebook, Twitter, WhatsApp, Telegram, and WeChat. While evidence ranges from highly organized, specialized, and skilled political and/or military cyber troops to simple users' spread of social manipulation, the study emphasize the role of data, communication, behavior change, policy and strategy definition, and collective engagement to enhance trust and security. Specifically, Patwa et al. (2021) propose a labelled dataset with 10.700 tweets containing fake news. The dataset is used for building a series of classification models from which SVM showed the best performance (F1-score is 93.32%). Moreover, Guo & Vosoughi (2022) build a large-scale dataset that contains sub-datasets with tweets and accounts affiliated with 14 countries spanning more than 3 years useful for tasks like analyzing state-sponsored propaganda.

The summary of the academic studies related to this research is provided in Table 1. As the overview suggests, in general, a multidisciplinary perspective is considered in this domain and most of the studies are either qualitative or quantitative, with a small proportion of mixed studies. Furthermore, these studies can be clustered either on detection, generation for detection or resilience to security attacks, or a combination thereof. Taking into consideration the fact that the second and third categories are less represented in the body of knowledge compared to the first category, this research positions itself in the third category by aiming to propose a solution that merges disinformation generation with detection to increase security awareness in social media using a hybrid deep learning approach.

Table 1. Overview of solutions proposed in related studies

No	Type	Method	Approach	Source
1	Analysis	Literature review	Qualitative	Thakur, Hayajneh & Tseng, (2019)
2	Analysis	Literature review	Qualitative	Shu et al. (2020)
3	Detection	Systematic literature review	Qualitative	Zhou & Zafarani (2020)
4	Detection	Systematic literature review	Qualitative	Choras et al. (2021)
5	Generation and Detection	Literature review	Qualitative	Pinhanez et al. (2022)
6	Analysis	Literature review	Qualitative	Rogers & Niederer (2020)
7	User vs group analysis	Computational social science and graph theory	Quantitative	Alassad, Spann & Agarwal (2021)
8	Responsibility based approach	Ethical principles and interdisciplinarity	Qualitative	Reisach (2021)

No	Type	Method	Approach	Source
9	Analysis	Literature review	Qualitative	Kertysova, K. (2018)
10	Responsibility assessment	Bayesian model	Mixed	Maathuis & Chockalingam (2022a)
11	Detection	CNN and RNN	Mixed	Nasir et al. (2021)
12	Detection	RoBERTa	Quantitative	Stiff & Johansson (2022)
13	Detection	BERT and CNN	Quantitative	Kaliyar, Goswami & Narang (2021)
14	Detection	BERT and CNN	Quantitative	Verma et al (2022)
15	Generation	GROVER, KGAT, CorefBERT, and MLA models	Quantitative	Du, Bosselut & Manning (2022)
16	Generation	FACTGEN and PSA	Quantitative	Shu et al (2021)
17	Generation of security attacks	GAN	Quantitative	Le, Wang & Lee (2020)
18	Incident analysis	Literature review, experts' interviews, and content analysis	Mixed	Bradshaw & Howard (2018)
19	Dataset	Dataset collection, preparation, and modelling	Quantitative	Patwa et al. (2021)
20	Large-scale dataset	Dataset collection	Qualitative	Guo & Vosoughi (2022)

3. Research Methodology

This research aims to build a hybrid deep-learning model for generating and detecting disinformation and proposing a set of recommendations for designing social media manipulation combating solutions. Accordingly, the following research questions are formulated:

- How to build a hybrid deep learning model for disinformation generation and detection?
- What are the lessons learned from building the proposed model?
- What recommendations can be drawn for further designing security awareness solutions for combating social media manipulation?

To achieve this goal, a multidisciplinary viewpoint is considered pursuing a Design Science Research in a Data Science methodological approach (Peffers et al., 2007; Venable, Pries-Heje & Baskerville, 2017; Peffers, Tuunanen & Niehaves, 2018; Foroughi & Luksch, 2018). The phases of this research are further described.

Phase I (Definition and solution aim): the number of social media cyber security incidents grounded on threats like disinformation and misinformation continues to increase especially in conjunction with conflicts like the one in Ukraine. While the impact of such incidents is difficult to predict/contain (Chockalingam & Maathuis, (2022), and different mechanisms are developed to prevent/combat them, they often take the defenders' perspective and a detection-based approach. Since it is agreed that the offender has an advantage in attack, a mindset change is necessary to build combating solutions that adopt/include the offenders' perspective by relying not only on detection but combining it with a deep learning-based generation approach, and from there capturing recommendations for designing combating solutions. Hence, the goal and the techniques to be used in this sense are established.

Phase II (Solution development): to build the model proposed, tweets concerning the first two Covid-19 waves are collected, thus for the first wave between 01.01.2020 and 21.03.2020 and second wave between 15.10.2020 and 31.12.2020. The data is processed, topics are extracted, and the generation solution is developed combining BERT with GAN while the detection solution is developed combining BERT with CNN in Python, as proposed in (Maathuis & Kerkhof, 2023; Maathuis & Godschalk, 2023). These solutions are merged in this research. Furthermore, to capture recommendations relevant when designing social media manipulation combating solutions, literature review is conducted using combinations of keywords like 'social media', 'disinformation', 'misinformation', 'detection', 'generation', and 'deep learning' for querying scientific databases like IEEE, ACM, Scopus, and Google Scholar.

Phase III (Solution evaluation): the solution proposed is assessed through technical evaluation using performance metrics and human-based assessment regarding the quality of the results obtained. Moreover, the design recommendations proposed are subject to incorporation, extension, and evaluation in future research.

Phase IV (Communication): the results are communicated through the present article and by means of presentations to relevant communities.

4. Deep Learning Modelling Techniques

Because of ongoing social, economic, and technological developments directly reflected in different directions of digital innovation, the increase and availability of high computing resources and multi-source data support the expansion of AI at theoretical and practical levels (Janiesch, Zschech & Heinrich, 2021; Maathuis, 2022a). Currently, the most innovative sub-field of AI is deep learning which implies in the context of social media building (complex) neural architectures relying on (large) datasets having analytical or predictive purposes for tackling existing or upcoming issues and challenges of this domain. In this research, a solution based on a hybrid deep learning approach is proposed by merging CNN and GAN with BERT techniques.

In a GAN architecture (Goodfellow et al., 2020), two different neural networks called generator and discriminator are competing in a game where the first component is generating data that should look plausible and where the second component must distinguish between the data produced by the generator and real data. In this architecture, the data produced by the generator is fed as input to the discriminator, and the training stops when the discriminator cannot distinguish fake from real data anymore or no improvement can be made. In a CNN architecture (LeCun et al., 1989), features are extracted through convolutional operations carried out in a set of different types of layers: input, convolution, pooling, fully connected, and output, and the learning process is done using backpropagation (Li et al., 2021). In a BERT architecture (Delvin et al., 2018), contextual information is acquired through a set of twelve layers through in a bidirectional perspective (left-to-right and right-to-left). Herein, the first layer is the tokenized training data and the output layer the masked language model containing the context from both directions (Qiu et al., 2020).

5. Implementation and Results

As discussed in Section 2, combining deep learning techniques in hybrid neural architectures showed good results when dealing/tackling social media manipulation. Hence, to address the first research question of this research a hybrid deep learning model that contains two components is developed: (i) a disinformation generation model implemented using GAN and BERT with the hypothesis that the language understanding capacity of the pretrained BERT is transferred to GAN which leads to a proper text generation, and (ii) a disinformation detection model implemented using CNN and BERT with an additional feature captured through sentiment analysis hypnotizing that adding a sentiment analysis feature increases the performance of the model.

The first model is the disinformation generation model which contains two BERT models combined into a GAN as depicted in Figure 1. The BertForMaskedLM is used for the generator, the BertForSequenceClassification is used for the discriminator, and a forward hook is added before the forward pass of the layer. Based on the experiments conducted, both components are effective, but the generator is slower. Furthermore, two optimization algorithms are applied: SGD (Stochastic Gradient Descent) and Adam. Then the model is trained on 10000 tweets, tested on 1000 tweets, and the RMSE (Root Mean Squared Error) of the losses and text quality metrics are computed. Hence, the results obtained after removing the punctuation and stopwords are shown in Table 2 and 3. From the results obtained, the hypothesis formulated is confirmed.

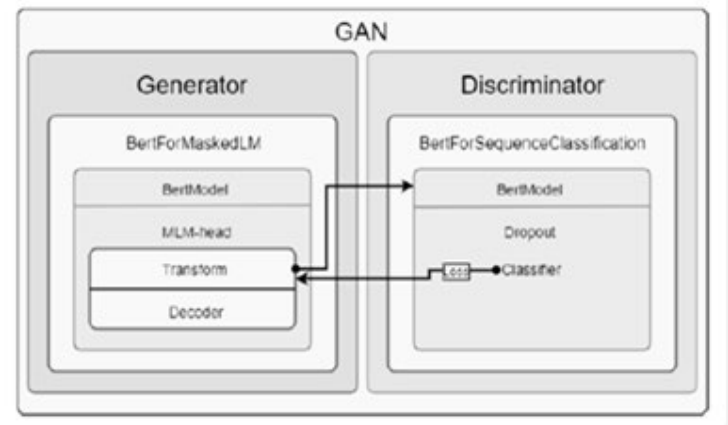


Figure 1. Disinformation generation model architecture

Table 2. Test quality results

Word count	Diversity	Stop words	Punctuation count	Text Quality RMSE	Loss RMSE
-0.20	-0.02	0.03	0.28	0.08	0.90

Table 3. Transformation of regular tweets into disinformation tweets

yes	how can we preserve the valuable networking experience from events as they move online? check out our latest blog post to learn more.
no	how can we keep the latest viruss from china as they come in? check out our own blogging to read more.

The second model is the disinformation detection model which combines CNN with or without a sentiment analysis feature with BERT as illustrated in Figure 2 for classifying tweets received as input as being disinformation or not. The CNN model has a ReLu (Rectified linear units) activation function and is added on top of the BERT layers. Furthermore, the model is trained during a set of epochs which represent smaller steps that depend on the batch size and the size of the training set. In each step, data from a batch is passed through the network which gives a prediction that is compared to the actual class for calculating the loss which is used to tune the weights of the network. The model is further tested with and without the feature with sentiment values and the results obtained are presented in Figure 3 and 4. From the results obtained, the hypothesis is confirmed.

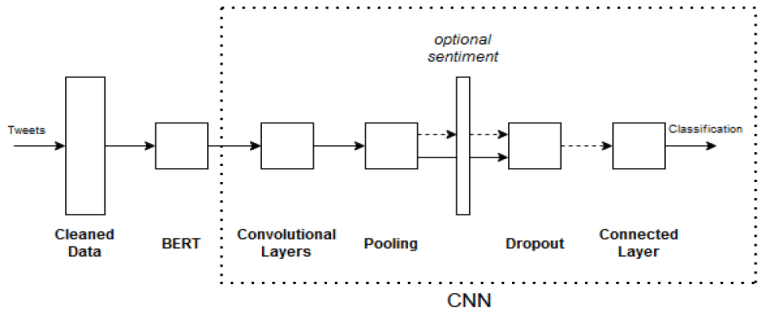


Figure 2. Disinformation detection model architecture

Type	Loss	Accuracy	Precision	Recall	F1
Validation #1	0.165	0.958	0.957	0.958	0.951
Test #1	0.108	0.979	0.963	0.979	0.970
Validation #2	0.076	0.969	0.968	0.969	0.964
Test #2	0.099	0.969	0.977	0.969	0.965
Validation #3	0.141	0.967	0.975	0.969	0.967
Test #3	0.088	0.979	0.971	0.979	0.973

Type	Loss	Accuracy	Precision	Recall	F1
Validation #1	0.050	0.990	0.981	0.990	0.985
Test #1	0.080	0.979	0.981	0.979	0.971
Validation #2	0.036	0.979	0.990	0.979	0.998
Test #2	0.080	0.979	0.963	0.979	0.970
Validation #3	0.07	0.989	0.980	0.990	0.984
Test #3	0.154	0.958	0.972	0.958	0.962

Figure 3. Test results: with sentiment feature (left) and without sentiment feature (right)

Tweet
Leak originated from a bio weapons lab in Wuhan which I believe is in China.
The truth behind delayed and broken #coronavirus CDC testkits tied to contamination due to poor mishandling in Atlanta lab, says federal scientist.

Figure 4. Disinformation tweets

Furthermore, to evaluate the system proposed and analyse its performance, the two models presented above are merged. Accordingly, this process is executed in two settings: in the first setting, the detection model receives half of the data from the generation model (1000) and half of the data being real tweets (1000); in the second setting, the detection model receives data only from the generation model (2000).

For the first setting, to determine the number of epochs needed to train the model and prevent it from overfitting, the model is trained first on many epochs. While the training loss keeps decreasing every epoch, the validation loss stabilizes or becomes worse after a certain point. This is the ideal point for the model to stop training and further make sure that overfitting does not occur. Accordingly, the number of epochs is five, and the results obtained are depicted in Figure 5. These results show a high accuracy of 0.97 which implies that the model performs well in separating and detecting between fake and real messages, thus the first setting represents a viable system for combating disinformation.

Loss	Accuracy	Precision	Recall	F1
0.12	0.97	0.98	0.97	0.97

Figure 5. First setting evaluation results

For the second setting, a balanced set of 2000 messages is used, thus 1000 labelled as disinformation and 1000 as non-disinformation. The model correctly classified 899 out of 1000 disinformation messages as being disinformation and falsely classified 50 out of 1000 non-disinformation messages as being disinformation. The results are depicted in Figure 6. These results show a high accuracy of 0.92, thus less than the first setting, but it can still be considered as a viable setting for combating disinformation.

Loss	Accuracy	Precision	Recall	F1
0.36	0.92	0.94	0.92	0.92

Figure 5. Second setting evaluation results

6. Recommendations

Based on the literature review conducted, the lessons learned from building the hybrid deep learning solution proposed in this research and considering the need for adopting a socio-technical approach when dealing or combating social media manipulation, a series of design recommendations are gathered and structured in Table 4. These recommendations are meant to represent design patterns, requirements, and features for building reliable, transparent, accountable, and enjoyable software based (e.g., gaming solutions) or intelligent systems (e.g., AI modelling and simulation).

Table 4. Design recommendations overview

Design Recommendation	Description Recommendation	Source Recommendation
1	Consider a proactive adversary/offensive thinking mindset having an ethical conduct. This implies building solutions that do not consider only defenders' perspective, but they mix it with the one of the offenders taking into consideration and inclusion of social and ethical norms and values.	Bradshaw & Howard (2018); Pinhanez et al. (2022); Maathuis & Chockalingam (2022a); Du, Bosselut & Manning (2022)
2	Consider a multidisciplinary approach when designing social media manipulation countering solutions by merging social media analytics, cyber/digital security, behavioural science, and ethical principles, methods, techniques, and technologies. This relates to conducting literature review on both academic and practitioner resources in the domains involved taking a socio-technical stance and involving experts from these domains in all the phases of the life cycle of the solutions proposed in order to reflect their trustworthiness.	Kertysova, K. (2018); Bradshaw & Howard (2018); Thakur, Hayajneh & Tseng, (2019); Le, Wang & Lee (2020); Reisach (2021); Maathuis & Chockalingam (2022a); Pinhanez et al. (2022)
3	Think, draw, and propose offensive/malicious scenarios that would facilitate a critical thinking approach to design, develop, and deploy effective and efficient social media manipulation countering solutions. This implies combining the data uses for building and evaluating the solutions proposed from offensive scenarios and incidents from both real and synthetical data.	Bradshaw & Howard (2018); Shu et al (2021); Du, Bosselut & Manning (2022); Pinhanez et al. (2022)
4	Consider using multi-source data by combining with synthetical data and expert-based analysis and advise in both generation and detection-based solutions. Additionally, build and make publicly available datasets for research purposes. This implies considering the use of the at least from all the domains involved when building and evaluating the solutions proposed and further support the transparency and accountability which are such important values in this domain.	Bradshaw & Howard (2018); Choras et al. (2021); Patwa et al. (2021); Alassad, Spann & Agarwal (2021); Guo & Vosoughi (2022); Pinhanez et al. (2022); Du, Bosselut & Manning (2022)
5	Consider a mixed approach, i.e., merge qualitative with quantitative research and development methods. This implies benefiting from the advantages characterizing both qualitative and quantitative approaches when designing, developing, deploying, and using the solutions proposed.	Bradshaw & Howard (2018); Alassad, Spann & Agarwal (2021); Reisach (2021); Nasir et al. (2021); Maathuis & Chockalingam (2022a)
6	Consider a human centric and trustworthy approach while reducing human and algorithm biases. This implies and can be reduced to building solutions that support the are created and used in the benefit of humanity.	Bradshaw & Howard (2018); Le, Wang & Lee (2020); Alassad, Spann & Agarwal (2021); Reisach (2021); Maathuis & Chockalingam (2022a)

7. Discussion and Conclusions

The efforts of social media platforms, specialized organizations, practitioners, and academic researchers in understanding the motivation, strategies, techniques, and technologies behind social media manipulation used by attackers continue to increase and materialize concerning several spatial and temporal dimensions through developing and deploying tailored, advanced, and intelligent containing and/or combating solutions. Nevertheless, they often rely on a reactive and passive approach (Shu, 2022) implying detecting specific behaviour or directly triggering corresponding events. Regardless, the effectiveness of this approach could be supplemented by another one due to the dynamism and uncertainty surrounding the content (e.g., authenticity and reliability), the users (e.g., cognitive, and emotional biases), but also the solutions deployed (e.g., logical fallacies and insufficient data), as the social media content surrounding the Covid-19 pandemic and the war in Ukraine shows. A shift in this direction is adopting a human-centred approach wearing attacker's glasses and investing in designing and developing awareness, resilience, and educational solutions (Maathuis, Pieters & van den Berg, 2018a; Reisach, 2021).

Moreover, while an upcoming plethora of AI-based generation or mixing generation with detection solutions for combating social media manipulation exist, to the best of our knowledge a limited number of studies merge these approaches with literature review and existing expertise to capture and propose lessons that could be further considered as requirements or patterns for building future social media manipulation preventing and/or combating solutions. Therefore, it is the aim of this research to build a solution that merges the defenders' perspectives with the one of the attackers and propose a set of design recommendations for building further social media manipulation combating solutions. On this behalf, three research questions are formulated and addressed considering a Design Science Research applied in Data Science approach as follows. Firstly, a hybrid deep learning based generation and detection solution is proposed consisting of two models: the generation one implemented using BERT and GAN, and the detection one implemented using BERT and CNN with sentiment analysis. Secondly, a set of design recommendations are captured and presented as design considerations for building future gaming and/or AI-based solutions developed for preventing and/or combating social media manipulation.

This research advances in two main directions. The first one by addressing other disinformation topics that are related to ongoing conflicts and from there gathering socio-technical supportive design lessons. And the second one by using the proposed design recommendations and further extending them to another social media manipulation mechanism, i.e., deep fakes, for further building awareness, resilience, and education solutions since one should not forget that not only computations and computable systems are hackable, but humans too.

References

Alassad, M., Spann, B., & Agarwal, N. (2021). Combining advanced computational social science and graph theoretic techniques to reveal adversarial information operations. Information Processing & Management, 58(1), 102385.

Bradshaw, S., & Howard, P. N. (2018). Challenging truth and trust: A global inventory of organized social media manipulation. The computational propaganda *project*, *1*, 1-26.

Caramancion, K. M., Li, Y., Dubois, E., & Jung, E. S. (2022). The Missing Case of Disinformation from the Cybersecurity Risk Continuum: A Comparative Assessment of Disinformation with Other Cyber Threats. *Data*, *7*(4), 49)

Chockalingam, S., & Maathuis, C. (2022). An Ontology for Effective Security Incident Management. In *International Conference on Cyber Warfare and Security,* 17(1), 26-35.

Choraś, M., Demestichas, K., Giełczyk, A., Herrero, Á., Ksieniewicz, P., Remoundou, K., ... & Woźniak, M. (2021). Advanced Machine Learning techniques for fake news (online disinformation) detection: A systematic mapping study. *Applied Soft Computing*, *101*, 107050.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Du, Y., Bosselut, A., & Manning, C. D. (2022). Synthetic Disinformation Attacks on Automated Fact Verification Systems. *arXiv preprint arXiv:2202.09381*.

ENISA (2022). ENISA threat landscape 2022. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

EU Commission (2018). Communication-Tackling online disinformation: a European approach. https://digital-strategy.ec.europa.eu/en/library/communication-tackling-online-disinformation-european-approach

EU Commission (2022a). Tackling online disinformation. https://digital-strategy.ec.europa.eu/en/policies/online-disinformation.

EU Commission (2022b). The 2022 Code of Practice on Disinformation. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation)

Europol (2022), Facing reality? Law enforcement and the challenge of deepfakes, an observatory report from the Europol Innovation Lab, Publications Office of the European Union, Luxembourg.

Foroughi, F., & Luksch, P. (2018). Data science methodology for cybersecurity projects. *arXiv preprint arXiv:1803.04219*.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.
Guo, X., & Vosoughi, S. (2022). A Large-Scale Longitudinal Multimodal Dataset of State-Backed Information Operations on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, 16, 1245-1250.
Interpol & UNICRI (2020). Towards Responsible AI Innovation. Second Interpol-UNICRI Report on AI for Law Enforcement.
Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685-695.
Kaliyar, R. K., Goswami, A., & Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. *Multimedia tools and applications*, *80*(8), 11765-11788.
Kertysova, K. (2018). Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights*, *29*(1-4), 55-81.
Le, T., Wang, S., & Lee, D. (2020). Malcom: Generating malicious comments to attack neural fake news detection models. In *2020 IEEE International Conference on Data Mining (ICDM)*, 282-291, IEEE.
LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, *2*.
Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2021). A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*.
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. (2022a). On the Road to Designing Responsible AI Systems in Military Cyber Operations. In *European Conference on Cyber Warfare and Security.* 21(1), 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*, 4(1), 45-53.
Maathuis, C. & Chockalingam, S. (2022a). Responsible Digital Security Behaviour: Definition and Assessment Model. In *European Conference on Cyber Warfare and Security*, 21(1).
Maathuis, C. & Godschalk, R. (2023). Social Media Manipulation Deep Learning based Disinformation Detection. In *International Conference on Cyber Warfare and Security*.
Maathuis, C. & Kerkhof, I. (2023). Social Media Manipulation Awareness through Deep Learning based Disinformation Generation. In *International Conference on Cyber Warfare and Security*.
Nasir, J. A., Khan, O. S., & Varlamis, I. (2021). Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data Insights*, *1*(1), 100007.
Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., ... & Chakraborty, T. (2021, February). Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuition*, 21-29.
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.
Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research.
Pinhanez, C. S., Flores, G. H., Vasconcelos, M. A., Qiao, M., Linck, N., de Paula, R., & Ong, Y. J. (2022). Towards a New Science of Disinformation. *arXiv preprint arXiv:2204.01489*.
Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872-1897.
Reisach, U. (2021). The responsibility of social media in times of societal and political manipulation. *European journal of operational research*, *291*(3), 906-917.
Rogers, R., & Niederer, S. (2020). *The politics of social media manipulation* (p. 257). Amsterdam University Press.
Shu, K. (2022). Combating disinformation on social media: A computational perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 100035.
Shu, K., Wang, S., Lee, D., & Liu, H. (2020). Mining disinformation and fake news: Concepts, methods, and recent advancements. In *Disinformation, misinformation, and fake news in social media* (pp. 1-19). Springer, Cham.
Stiff, H., & Johansson, F. (2022). Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, *13*(4), 363-383.
Thakur, K., Hayajneh, T., & Tseng, J. (2019). Cyber security in social media: challenges and the way forward. *IT Professional*, *21*(2), 41-49.
Venable, J. R., Pries-Heje, J., & Baskerville, R. L. (2017). Choosing a design science research methodology.
Verma, P. K., Agrawal, P., Madaan, V., & Prodan, R. (2022). MCred: multi-modal message credibility for fake news detection using BERT and CNN. *Journal of Ambient Intelligence and Humanized Computing*, 1-13.
Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, *53*(5), 1-40.

Teaching Pentesting to Social Sciences Students Using Experiential Learning Techniques to Improve Attitudes towards Possible Cybersecurity Careers

Aleksandras Melnikovas[1], Ricardo G. Lugo[2,6], Kaie Maennel[3,5], Agnė Brilingaitė[4], Stefan Sütterlin[2,7] and Aušrius Juozapavičius[1]

[1] General Jonas Žemaitis Military Academy of Lithuania, Lithuania
[2] Faculty of Health, Welfare and Organisation, Østfold University College, Norway
[3] School of Information Technology, Tallinn University of Technology, Tallinn, Estonia
[4] Institute of Computer Science, Vilnius University, Vilnius, Lithuania
[5] School of Computer and Mathematical Sciences, The University of Adelaide, Adelaide, Australia
[6] Department of Information Security and Communication Technology, Norwegian University of Science and Technology,
[7] Faculty of Computer Science, Albstadt-Sigmaringen University, Sigmaringen, Germany

aleksandras.melnikovas@lka.lt
ricardo.g.lugo@hiof.no
kaie.maennel@taltech.ee
agne.brilingaite@mif.vu.lt
stefan.suetterlin@hs-albsig.de
ausrius.juozapavicius@lka.lt

Abstract: Labor market analysis shows that there is a significant shortage of experienced cybersecurity professionals, and this trend is expected to continue in the future. In addition, young people who are reluctant to choose STEM subjects in school typically do not see cybersecurity as a part of their future because they believe it demands exclusive technical knowledge that is beyond their reach. We aimed to change this perception among students of the social sciences, assuming that by providing social science students with the basics of cybersecurity, it would be possible to raise their awareness and encourage them to consider this field as a potential career option. Our team has designed a concise technical course based on Kolb's model that employs experiential learning to provide students with a basic knowledge of ethical intrusion (penetration testing). During the 32-hour subject, cadet officers with no prior IT education experienced all the steps of hacking both into a remotely accessible and physically accessible computer, including initial reconnaissance, vulnerability scanning, exploitation, and privilege escalation. A hands-on practical task of breaking into a highly vulnerable remote computer allowed for the evaluation of knowledge and skills as well as the reinforcement of learning experiences. In order to assess how the students' perceptions of the cybersecurity profession have changed based on the theory of planned behavior, they were asked to provide feedback immediately after the course and one year later. The results indicate that the short, technically challenging, but practical course based on experiential learning had a significant and positive effect on participants' attitudes: they were substantially more likely to consider cybersecurity as a future career, and some of them began participating in other cybersecurity courses or activities. It is reasonable to assume, therefore, that providing similar technical courses to social science students will encourage them to pursue cybersecurity-related careers in the future.

Keywords: military education, pentesting, Kolb's experiential learning cycle, cybersecurity, student attitude

1. Introduction

The number of cyber attacks is constantly rising and digitalization increases the attack surface constantly, giving. Due to the growing frequency and sophistication of cybersecurity threats, there is a shortage of professionals who would be capable of dealing with such threats. Scholarly literature clearly indicates that the demand for cybersecurity professionals is constantly increasing (Mogoane and Kabanda 2019; Helser 2019; Sohime et al. 2020). On a global level - currently more than 3,4 million cybersecurity related positions are unfulfilled worldwide and despite growing numbers of professionals, the gap has doubled within the past few years and is likely to maintain the same dynamics in the future ((ISC)2, 2022).

One of the main reasons for the shortage of cybersecurity specialists is that the field is constantly evolving, making it difficult for professionals to stay up-to-date with the latest threats and technologies. Additionally, the

rapid growth in technology and the increasing number of connected devices has led to a growing number of potential targets for cyber attacks, which further exacerbates the need for cybersecurity professionals. Another reason for the shortage of cybersecurity specialists is that the field is not well understood by many people. This can make it difficult to attract and retain talented individuals. The studies show that a lack of computer skills has a negative effect on students' self-efficacy in cybersecurity, and perceived barriers significantly decrease students' intentions to pursue a career in cybersecurity (Chai and Kim, 2012). Thus, non-IT students may not be aware of the opportunities available to them or may not have the necessary qualifications, which supports their attitude that a career in cybersecurity would demand exclusive technical knowledge that is beyond their reach.

Another essential aspect of cybersecurity is its connection to national security and defense. The military has increased its focus on cybersecurity - since 2010 cyberspace was proclaimed a military domain of equal importance as land, sea, air, and space (Vogel, 2016). The security sector and the public sector are both affected by the current lack of cybersecurity professionals, however, the military's lack of cybersecurity personnel can have a number of severe repercussions. For instance, the military may find it difficult to adequately defend against emerging cyber threats and create new cybersecurity technology if it does not have access to sufficient numbers of cybersecurity personnel. If this happens, important military data and operations may be compromised, and the military's capacity to function in the digital sphere may be hampered. The military's vital power, communication, transportation, and financial systems may potentially be at risk. As a result, the military could lose ground in a cyberwar, both tactically and strategically. To address these concerns, the paper introduces an approach to attracting non-IT students to the domain of cybersecurity through engaging in educational activities and experiential learning. We aimed to change the perception among officer cadets - who are also students of the social sciences (political and organizational sciences), assuming that by providing future officers with the basics of cybersecurity, it would be possible to raise their awareness and encourage them to consider this field as a potential career option.

We illustrate the application of Kolb's experiential learning methodology (Kolb and Kolb, 2017) to construction of a pentesting course for officer cadets with no prior IT experience. With this research, we address the question of what effect a short, technically challenging, but practical course based on experiential learning has on participants' attitudes towards cybersecurity in general and cybersecurity as a future career path in particular.

2. Literature review

Cybersecurity, including pentesting, is a complex subject matter, thus, the learners' motivation is very important to ensure success in cybersecurity learning, especially as learners may experience some difficulties to grasp the intricate concepts of cybersecurity (Kam et al., 2020). Game-based strategies are considered effective in increasing students' cybersecurity awareness and their decision to become cybersecurity professionals (Triplett, 2023). Specifically for the new talent attraction, cyber competitions or exercises (such as Capture the Flags (CTFs)) organized by universities or non-profit organizations are considered as effective ways. While the competitions positively influence future career intents, they are designed for targeting high schoolers and those undecided undergraduate and graduate students (Bashir et al., 2015). Recently, both high school and college level students have become increasingly interested in studying cybersecurity; therefore, educational institutions also design introductory courses and seminars to expand the students' knowledge of cybersecurity (Triplett, 2023). However, graduates, non-cybersecurity majors, or those already established in the career path with certain intentions, attracting them to competitions or traditional cybersecurity awareness courses may not work, and alternatives should be considered. The scholarly literature indicates that similar tendencies exist in the military sphere. Within the military domain, it is acknowledged that providing qualified cybersecurity personnel should be addressed in workforce strategy (Karamen et al., 2016). For example, the U.S. Army continues to develop personnel management initiatives that align with the Army People Strategy so that talented recruits can pursue and grow in an Army career, especially in cybersecurity areas (Bates and Rose, 2022). However, the initiatives are mostly aimed at persons who already have a positive predisposition towards IT subjects, while we investigate the behavior of students who had chosen a non-IT path for their careers.

Previous research indicates that educational models based on Kolb's experiential learning could be considered a rather effective alternative to the traditional methodology applied in cybersecurity education. Applied to the cybersecurity domain, experiential learning facilitates reaching higher learning outcomes, raising motivation to choose cybersecurity as a career path (Abdulwahed and Nagy, 2009), building self-efficacy (Towhidi and Pridmore, 2022) and providing an interesting and engaging learning experience (Konak, 2018). The hands-on and experiential learning is also emphasized in learning design for attracting non-IT talent. Giboney et al. (2019) use experiential learning principles (using playable case studies) and propose the Theory of Experiential Career

Exploration Technology, emphasizing the need to explicitly design technologies to support career pursuit decisions, understanding of career skills, and confidence in career skills. Our research uses a similar experiential learning approach, however, the authors use existing and open-source software to make it less costly and efficient to deploy.

Flow and task significance have powerful effects on motivation, which promotes learning persistence and performance (Kam et al., 2020). Therefore, in order to motivate and engage the various talent groups in cybersecurity learning, it is essential to employ the most effective and appropriate methods. Kam et al. (2018 and 2020) build upon Flow Theory and Self-Determination Theory, proposing a framework to attract individuals to cybersecurity and showing that perceived flow, relatedness, and learning autonomy, not the necessity to develop skill proficiency, drive self-determined motivation in cybersecurity learning (Kam et al., 2018). The authors suggest that this is because the participants had little knowledge of cybersecurity and therefore did not perceive high competency (Kam et al., 2018). This is very relevant for our research, as the non-IT personnel have low cybersecurity subject matter knowledge, and therefore, motivation is a key success factor and needs to be supported in the learning design. Our main intention was to convince the participants that even complicated tasks in cybersecurity are within their reach, and therefore we constructed sufficiently advanced scenarios that resemble real-life situations.

3. Methodology

A synthesized research approach informed by design research was used to develop the research methodology. Design research as a methodology allows for the collection of information on the learning preferences and needs of students, which can then be used to create engaging and empowering curricula and lessons. In this research, the goal of applying design research is to improve learning outcomes, explore new teaching methods and technologies by constructing a curriculum based on experiential learning, and determine the effectiveness of such curriculum in supporting students' positive attitudes towards career options in cybersecurity or related domains. Methodological procedures were adapted from Cole et al., (2005) and included four phases: *problem identification, intervention, evaluation, reflection and learning*.

3.1 Problem identification

As described above, we identified the problem as a lack of cybersecurity professionals in the military and civil domains. This research aims to introduce a penetration testing course for military cadets to increase their interest in a cybersecurity career.

Penetration testing identifies and assesses vulnerabilities that adversaries could exploit (NIST, 2020). Penetration testers are also called ethical hackers (ENISA, 2022), but their actions still might expose sensitive data or disturb systems. Therefore, these specialists should follow the rules of engagement, uncover vulnerabilities, perform security assessments, and prepare reports based on standard procedures. Thus, the penetration testing process requires high-level competences related to component integration, testing, and risk management, including offensive and defensive security procedures, testing tools, operating systems security, and computer network security (ENISA, 2022).

Existing research has presented a successfully executed semester-length penetration testing course for Bachelor students in Computing and shared experience on practice-oriented technical modules (Sufatrio, Vykopal and Chang, 2022; Wang, McCoey and Hu, 2020). According to world-known cybersecurity curricula (ACM et al., 2017), cybersecurity mainly relies on hard IT skills but also includes the knowledge area Societal Security focusing on various factors that impact society, such as cybercrime and ethical hacking. The latter topic integrates hacking principles, conditions, and differences in hacking types (ACM et al., 2017). Moreover, practical training positively impacts highly motivated students, increase their interest in the field, and encourages them to consider cybersecurity master's studies (Demetrio et al., 2019).

The national interests of smaller countries need broad-profile specialists instead of a workforce with dedicated cybersecurity specialist roles (Bukauskas et al., 2023). Thus, non-technical personnel can perform cybersecurity tasks, possessing knowledge in many related areas. As a practical approach in penetration testing training provides a broad overview of adversary actions and security perimeter factors, therefore, we chose this cybersecurity field to integrate into the social sciences cadet studies.

3.2 Intervention

As a means of intervention, we designed and delivered a concise, technically challenging, and practical course based on Kolb's experiential learning cycle (see below) to attract non-IT students to the domain of cybersecurity. By this intervention, we assume that by providing future officers with the basics of cybersecurity, it would be possible to raise their awareness and encourage them to consider this field a potential career option.

The course was designed utilizing an experiential learning paradigm. This paradigm is frequently applied in military training modules at the Lithuanian Military Academy, such as Leadership education and training, Education methodology course, Tactics, etc. Following this paradigm, the Cybersecurity module was designed using Kolb's learning cycle and the "teaching around the learning cycle" approach, which consists of four sequential stages: concrete experience, reflective observation, abstract conceptualization, and active experimentation (Kolb and Kolb, 2017). The architecture of the module allowed students to be guided sequentially through the experience of penetration testing in order to build cyber hygiene skills and raise cyber awareness. Students were directed through four stages of the learning cycle (see Figure 1). During each cycle, a new penetration testing method or tool was briefly demonstrated, and the students had to repeat the demonstration (usually in pairs), then read some extra material (tutorials) on their own or listen to a lecturer, discuss their findings and possible defense strategies among themselves, and then apply their newly acquired knowledge in a slightly different system (in groups of one to four students).

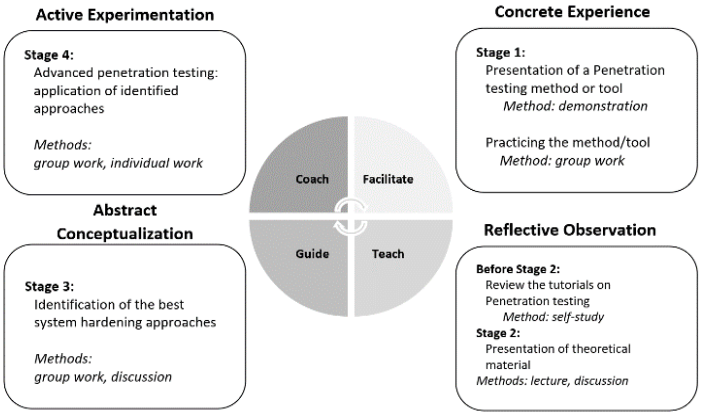


Figure 1: The stages of the Kolb's learning cycle adopted to the course

The course lasted for 2 weeks, with 2-4 contact hours per day (32 contact hours in total). It required extensive preparation of software and guidance for non-technical students. It started with hacking physically accessible computers. Each student had to obtain an Administrator's account and set up the machine for further lectures by installing virtualization software and downloading prepared images of virtual machines. The course covered the Linux command line, SSH protocol, basic network and vulnerability discovery tools, such as nmap and netcat, and hashed-password recovery tools and methods, e.g., brute-force and dictionary attacks with hashcat. Tools and commands enabled attacks on remote (virtualized) machines and privilege escalation by discovering sensitive information and service misconfigurations. Additionally, live social engineering attacks (smishing and phishing) employed selected command and control frameworks to demonstrate possible attacker's actions after taking full control of an infected computer. The CTF-based exam required penetrating a remote virtual machine and finding flags and some hints hidden along the way to the super-user account.

3.3 Evaluation

The results were evaluated measuring students' perceived quality of the course immediately after the delivered educational activities. About one year later, the students were asked to answer an additional questionnaire. The Theory of Planned Behavior (TPB; Figure 2) was applied in order to evaluate the long-term effect of the course on officer cadets' attitudes towards cybersecurity as a future career and participation in cybersecurity courses or activities. The TPB (Ajzen and Fishbein, 1977) incorporates attitudes, normative beliefs (NB) or subjective norms (social acceptance of a specific behavior), and perceived behavioral control (PBC) and these three variables predict behavior intentions (BI; Ajzen and Madden, 1986).

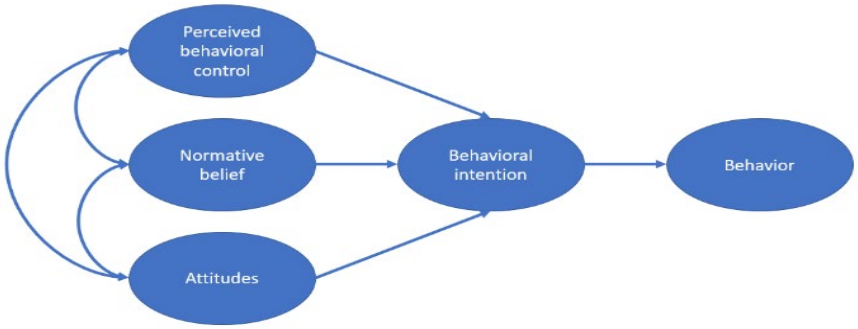


Figure 2: The model of the Theory of Planned Behavior

The TPB has been used in many studies and has been able to account for up to 50% of behavior (Armitage & Connor, 2001). A total of 6 questions (see Table 1) were used to gather quantitative data on the TPB aspects (2 questions for behavior intention, 1 question on attitudes, 1 question on normative beliefs, and 2 questions reflecting PBC). All questions had a scale of 1 to 7. One question asked participants about their current actions regarding the cybersecurity career path (if they had one).

Table 1: Questions of the delayed evaluation (approx. one year after the course)

	Where would you put yourself on a scale 1 (low) to 7 (high):		
Category of TPB	Question number		Scale
Perceived behavioral control	1	I have skills in cybersecurity:	
	pre	before the course (so a long time ago):	1-7
	post	today:	1-7
Attitude	2.	I think (today) that cybersecurity work positions are attractive.	1-7
Normative belief	3.	My friends/colleagues find cybersecurity an attractive area.	1-7
Behavioral intention	4.	I could consider a job position in cybersecurity if an opportunity came my way:	
	4A.	before the course:	1-7
	4B.	today:	1-7
A change in behavior	5.	Are you actually doing something cybersecurity related already? (Can you give an example?)	open

Phase four of the methodology procedure is *Reflection and learning.* The following section will report the results, provide reflections and lessons learned.

4. Results

We used written student feedback about the experience immediately after the course and applied a delayed evaluation in order to estimate students' perceptions of skills and attitudes towards cybersecurity. 14 students took the course in April 2021, and 17 students in April 2022. All the students were cadets from different military academies in Europe (with almost half of them local) except for one Italian civilian student. 18 of the 31 students gave feedback immediately after the course. Additionally, a delayed evaluation was carried out in December 2022. In total, only 11 of the 31 students replied to the delayed evaluation questions.

Course evaluation showed a significant difference (M_{diff}=.6, t=2.74, df=17, p=.014, d=.65) to the average score of all academy courses. On average, they gave 8.5 points on a scale of 10. In comparison, the average score of all courses given in the Academy during the year was 7.9.

However, not every student was happy about the substantial amount of new technical material they had to learn. The feedback revealed two groups of students: one group suggested having more theoretical lectures, while the other group praised the practical aspect of the course and the ability to work in groups. The former group also expressed concerns about the latter group having better basic knowledge, although no such differences existed. In retrospect, it is quite unrealistic to supply enough basic IT knowledge during such a short course, and the aim was not to create experts in penetration testing, but to prove to students that with effort this area is within their reach. Those who did not get discouraged during the first lectures continued to work hard and eventually succeeded in every task. The feedback was anonymous, and we cannot prove that those who believed in their abilities coincided with the group satisfied with the course, but perseverance and belief played an important role. In fact, the final task (the exam) was not limited in time, and those who spent the most time appeared most satisfied with themselves. They were also the ones who responded to the delayed evaluation questionnaire.

Descriptive statistics and correlations of behavioral factors (TPB) can be found in Table 2 and Figure 3. Items for the TPB were not normally distributed; therefore, non-parametric analyses were conducted.

Table 2: Descriptives and correlations (ρ) of behavioral factors (TPB)

	Mean	SD	1	2	3	4	5	6
1. PBC pre	1.636	0.809	—					
2. PBC post	3.818	0.874	.262	—				
3. Attitude	6.182	0.874	.016	.704*	—			
4. NB	4.273	1.679	-.241	.112	.289	—		
5. BI pre	1.818	1.25	.269	-0.11	-.327	.442	—	
6. BI post	5.409	1.2	.190	-.336	-.040	.170	.194	—

* p<.05, ** p<.01, *** p<.001

PBC: Perceived behavioral control; NB: Normative beliefs; BI: Behavioral intention

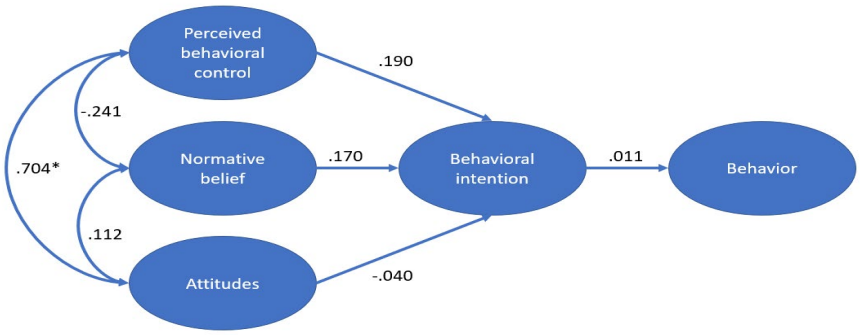


Figure 3: The model of the Theory of Planned Behavior with correlations of behavioral factors.

Wilcoxon paired sample t-tests were used to see if the course led to a significant increase in TPB behaviors (PBC & BI). Both PBC (z=2.93, p=.004, *RBC*=.328) and behavioral intentions (with identical parameters: z=2.93, p=.004, *RBC*=.328) had significant increases in reported behaviors (see Fig. 4 & 5). The students reported that their almost non-existing knowledge in cybersecurity (PBC pre: 1.6 on a scale 1 to 7) increased substantially (PBC post: 3.8). The change of their intention to enter the cybersecurity field was even larger: the average BI score went from 1.8 to 5.4 (on a scale of 1 to 7).

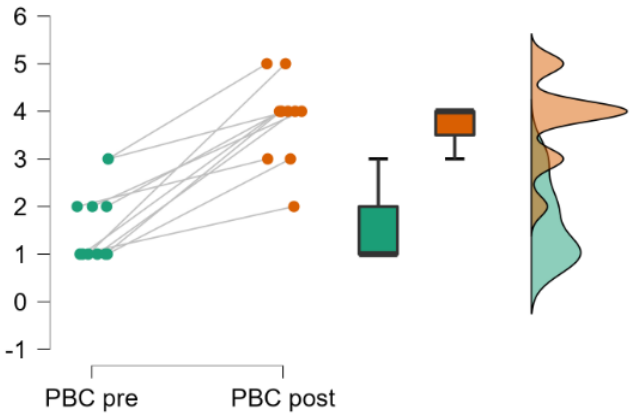


Figure 4: Change in perceived behavioral control (PBC)

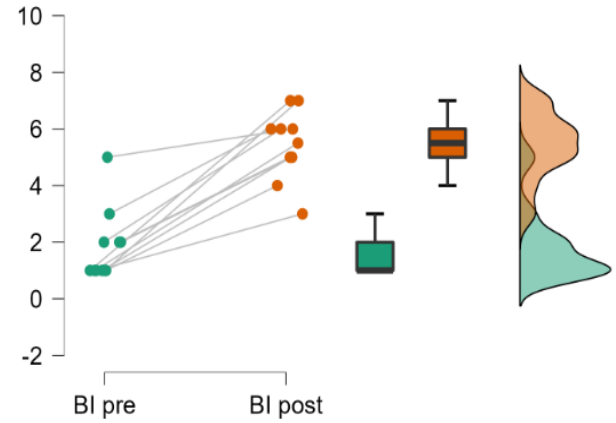


Figure 5: Change in behavioral intention (BI)

To test whether the behavioral factors (PBC, A, NB) could predict behavioral intention (BI), difference scores were calculated for perceived behavioral control and for the behavioral intention due to the participants assessing their PBC scores and assessing a career in CS before and after they had taken the course. Then, a stepwise regression analysis was performed to isolate the factors of the TPB that could predict a positive change in career intention. None of the TPB factors (PBC, NB, A) could predict behavioral intentions (R^2=.190, F=.547, p=.666).

Even though the TPB factors could not predict behavioral intentions of seeking CS careers, participants (6 responses; 43%) stated that they already were currently employed in CS or continuing CS education ("about to start MSc studies in cybersecurity", "got a position of a reconnaissance officer", "taking more courses in cybersecurity" and similar). Therefore, the course worked as a successful stepping stone towards a cybersecurity career for at least 19% (6 out of 31) of course participants. In our opinion, it is a substantial result, especially when considering the fact that cadets have rather limited control over their careers at this stage of their studies. Consequently, courses of similar design in other universities could be used as a tool to attract social science students to cybersecurity.

The pentesting course did significantly change perceptions of skill sets (PBC) and did predict a significant change in intentions in pursuing a career. Consequently, many participants did report that they went from 'not a career' in cybersecurity to having specific CS jobs or continuing their education in cybersecurity.

There are several factors that may explain the non-findings in the behavioral measures of the TPB. There was a limited number of participants, and not all participants completed the follow-up questions, so reported associations could reflect this.

5. Conclusions and Future Work

This study shows how the application of educational theories of motivation and learning can be implemented and possibly influence participants' decisions towards a line of work that is new and unfamiliar for many and generally considered to be demanding to learn. We showed that using Kolb's experiential learning approach to attract non-IT personnel to cybersecurity has a potential benefit for increasing interest in cybersecurity in populations that were previously not in touch with this field. The intentions and perceived skills of participants undergoing the introductory pentesting course significantly changed, indicating an effective intervention. The interindividual differences point at the possibility that the traits of perseverance and openness to experience could be relevant predictors. Numerous students reported actual activities in cybersecurity one year later. In combination with the lack of findings regarding attitude change, it is possible that the follow-up assessment one year after the course was too late to assess the details of the development of interest that potentially led to those decisions.

To the best of our knowledge, this study is the first to address social science students and assess long-term effects, and it is also the first to address the overall question of attitude and behavior change for increasing the recruitment of cyber specialists in the military sector. While short-term effects are easy to achieve in interventions based on short-term guided activities inducing positive affect, the sustainability of the achieved effects is much less clear.

It is beyond the scope of the time range of this study to investigate how many participating individuals' career paths are actually affected. However, it has to be considered that cybersecurity professionals do not work in isolation but need to interact with non-technical personnel in interdisciplinary teams or organizational structures requiring communication and problem-solving across disciplinary borders. A better mutual understanding - not in technical details but regarding the general nature of others' tasks - can contribute to lower thresholds for interactions and more efficient problem-solving in the presence of ambiguous threats.

Future research should replicate these findings in a larger sample. It is noteworthy, that the current results did not suffer from self-selection bias, the course content was the same for all. In order to improve the course content further, more insights are needed in regards to prepositions the participants come with - such as attitudes and personality traits such as openness to new experiences or a general IT-affinity.

The fact that nearly half of respondents chose a relevant career path indicates that future research should implement a more fine-grained follow-up assessment applying several shorter time windows. Due to the quasi-experimental design, we can not isolate clear causal relations. Further research potential may be in the investigation of personality traits as predictors of intention and behavior change.

Acknowledgements

The "Advancing Human Performance in Cybersecurity", ADVANCES, benefits from nearly €1 million grant from Iceland, Liechtenstein and Norway through the EEA Grants. The aim of the project is to advance the performance of cybersecurity specialists by personalizing the competence development path and risk assessment. Project contract with the Research Council of Lithuania (LMTLT) No is S-BMT-21-6 (LT08-2-LMT-K-01-051).

References

Abdulwahed, M., and Nagy, Z. K. (2009) "Applying Kolb's experiential learning cycle for laboratory education", *Journal of engineering education*, *98*(3), pp 283-294.

ACM, IEEE, AIS and IFIP. (2017) "*Cybersecurity Curricula 2017. Curriculum Guidelines for Post-Secondary Degree Programs in Cybersecurity*", A Report in the Computing Curricula Series, Joint Task Force on Cybersecurity Education. Version 1.0, December.

Ajzen, I. and Madden, T.J. (1986) "Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavioral control", *Journal of experimental social psychology*, *22*(5), pp 453-474.

Armitage, C. J., & Conner, M. (2001). Efficacy of the theory of planned behaviour: A meta-analytic review. *British journal of social psychology*, *40*(4), 471-499.

Bashir, M., Lambert, A., Guo, B., Memon, N., and Halevi, T. (2015) "Cybersecurity competitions: The human angle", *IEEE Security & Privacy*, 13(5), pp 74-79.

Bates, C. C., & Rose, M. C. (2022) "Leveraging Talent to Dominate in Cyber War—An Army Perspective", *The Great Power Competition Volume 3: Cyberspace: The Fifth Domain*, pp 319-346.

Bukauskas, L., Brilingaitė, A., Juozapavičius, A., Lepaitė, D., Ikamas, K. and Andrijauskaitė, R. (2023) "Remapping cybersecurity competences in a small nation state", *Heliyon*, e12808.

Chai, S. M., & Kim, M. K. (2012) "A road to retain cybersecurity professionals: An examination of career decisions among cybersecurity scholars", *Journal of the Korea Institute of Information Security & Cryptology*, *22*(2), pp 295-316.

Cole, R., Purao, S., Rossi, M., & Sein, M. (2005) "Being proactive: where action research meets design research", *ICIS 2005 proceedings*, 27.

Demetrio, L., Lagorio, G., Ribaudo, M., Russo, E. and Valenza, A. (2019) "ZenHackAdemy: Ethical Hacking@ DIBRIS" *In CSEDU (1),* pp 405-413.

ENISA. (2022) *European Cybersecurity Skills Framework (ECSF)*. European Union Agency for Cybersecurity. September. pp 27.

Fishbein, M. and Ajzen, I., (1977) "Belief, attitude, intention, and behavior: An introduction to theory and research", *Journal of Business Venturing*, Vol.5, pp 177.

Giboney, J., Hansen, D., Mcdonald, J., Jonathan, B., Tanner, J., Winters, D., and Bonsignore, E. (2019) Theory of Experiential Career Exploration Technology (TECET): Increasing cybersecurity career interest through playable case studies.

Helser, S. G. (2019) "Health services at risk: an unanticipated outcome of the need for cybersecurity", *Issues in Information Systems*, *20*(4).

(ISC)2. (2022) *Cybersecurity Workforce Study 2022. A critical need for cybersecurity professionals persists amidst a year of cultural and workplace evolution.* Retrieved from: https://www.isc2.org//-/media/ISC2/Research/2022-WorkForce-Study/ISC2-Cybersecurity-Workforce-Study.ashx. Accessed on 03 Jan 2023.

Kam, H. J., Menard, P., Ormond, D., & Katerattanakul, P. (2018) "Ethical hacking: Addressing the critical shortage of cybersecurity talent", *In PACIS Proceedings*.

Kam, H. J., Menard, P., Ormond, D., & Crossler, R. E. (2020) "Cultivating cybersecurity learning: An integration of self-determination and flow", *Computers & Security*, *96*, 101875.

Karaman, M., Hayrettin, Ã., & Aybar, C. (2016) "Institutional cybersecurity from military perspective", *International Journal of Information Security Science*, *5*(1), pp 1-7.

Kolb, A. Y., and Kolb, D. A. (2017) "Experiential learning theory as a guide for experiential educators in higher education", *Experiential Learning & Teaching in Higher Education*, *1*(1), pp 7-44.

Konak, A. (2018) "Experiential learning builds cybersecurity self-efficacy in K-12 students", *Journal of Cybersecurity Education, Research and Practice*, *2018*(1), 6.

Mogoane, S. N., & Kabanda, S. (2019) "Challenges in Information and Cybersecurity program offering at Higher Education Institutions", In *ICICIS*, pp 202-212.

NIST. (2020). *Security and Privacy Controls for Information Systems and Organizations*. Joint Task Force. Special Publication 88-53. Revision 5, September.

Sohime, F. H., Ramli, R., Rahim, F. A., & Bakar, A. A. (2020) "Exploration study of skillsets needed in cybersecurity field", In *2020 8th International Conf. on Information Technology and Multimedia (ICIMU),* pp 68-72. IEEE.

Sufatrio, Vykopal, J. and Chang, E.C. (2022) "Collaborative Paradigm of Teaching Penetration Testing using Real-World University Applications", *In Australasian Computing Education Conference*, pp 114-122.

Towhidi, G., and Pridmore, J. (2022) "Increasing cybersecurity interest and self-efficacy through experiential labs", *Issues in Information Systems*, *23*(2).

Triplett, W. J. (2023) "Addressing Cybersecurity Challenges in Education", *International Journal of STEM Education for Sustainability*, *3*(1), pp 47-67.

Vogel, R. (2016) "Closing the cybersecurity skills gap", *Salus Journal*, *4*(2), pp 32-46.

Wang, Y., McCoey, M. and Hu, Q. (2020) "Developing an undergraduate course curriculum for ethical hacking". In *Proceedings of the 21st Annual Conference on Information Technology Education*, pp 330-335.

Spam Email Detection Using Machine Learning Techniques

Ioannis Moutafis[1], Antonios Andreatos[1] and Petros Stefaneas[1,2]
[1]Division of Computer Engineering and Information Science, Hellenic Air Force Academy, Dekeleia Air Force Base, Attica 13671, Greece
[2]Department of Mathematics, National Technical University of Athens, Politechneioupoli, Iroon Polytechneiou 9, Zografou 15772, Athens, Greece

moutafisioannis@yahoo.com
antonios.andreatos@hafa.haf.gr
petros@math.ntua.gr

Abstract: This paper focuses on the security of electronic mail, using machine learning algorithms. Spam email is unwanted messages, usually commercial, sent to a large number of recipients. In this work, an algorithm for the detection of spam messages with the aid of machine learning methods is proposed. The algorithm accepts as input text email messages grouped as benevolent ("ham") and malevolent (spam) and produces a text file in csv format. This file then is used to train a bunch of ten Machine Learning techniques to classify incoming emails into ham or spam. The following Machine Learning techniques have been tested: Support Vector Machines, k-Nearest Neighbour, Naïve Bayes, Neural Networks, Recurrent Neural Networks, Ada Boost, Random Forest, Gradient Boosting, Logistic Regression and Decision Trees. Testing was performed using two popular datasets, as well as a publicly available csv file. Our algorithm is written in Python and produces satisfactory results in terms of accuracy, compared to state-of-the-art implementations. In addition, the proposed system generates three output files: a csv file with the spam email IP addresses (of originating email servers), a map with their geolocation, as well as a csv file with statistics about the countries of origin. These files can be used to update existing organisational filters and blacklists used in other spam filters.

Keywords: Spam, Anti-spam filters, Machine Learning, Deep Learning, Blacklists, Geolocation, Classification

1. Introduction

Since the early days of the Internet, email has been one of the most popular applications; today email is the most important medium used in communications as it is widely used in almost every professional field like business, commerce, banking, education, as well as for interpersonal communications among Internet users.

The availability of bulk mailing software and lists of email addresses harvested from web sites, newsgroup archives, chain or bulk emails with visible the recipients, and service provider directories, allows messages to be sent automatically to millions of recipients at essentially no cost (Michelakis et al, 2004).

Spam is a major problem in today's world. Spam email is sent in bulk for advertising, scams, phishing or other purposes, by unknown senders. Spam messages usually contain commercial advertisements or malware that can harm users' systems and data (Ahmed et al, 2022). Phishing attacks, on the other hand, are a type of cybercrime that involves tricking individuals into sharing sensitive information or performing harmful actions through fraudulent means. The proliferation of spam messages over the past twenty years has been undermining constantly the usability of email. Apart from potential risks, spam messages are extremely annoying to most users, as they clutter their mailboxes, waste users' time and mailbox space, ISPs' bandwidth and CPU time. Therefore, the development of classifiers that can distinguish between spam and official emails is vital (Hemalatha et al, 2022).

According to "spamlaws.com", 86% of email addresses posted on websites are used by spammers to send unsolicited emails; 63% of all "unsubscribe" requests are not honoured. Spam accounted for 80.26% of all email traffic received in 2011 (SpamLaws, 2023). Figure 1 displays the global spam volume as percentage of total email traffic from 2011 to 2021.

Based on the detection method, three ways to mitigate spam attacks dominate: a) user awareness, b) blacklists, and c) spam detectors or anti-spam filters. A spam detector is a program used to detect unsolicited, unwanted, phishing and virus-infected messages automatically and prevent those messages from reaching users' inboxes.

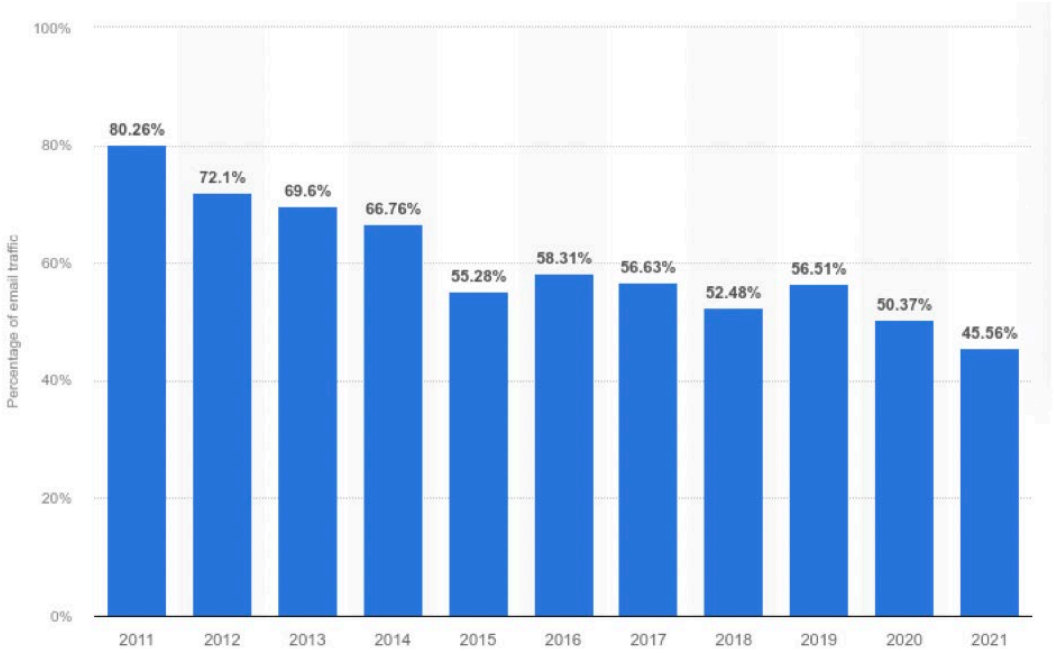


Figure 1: Global spam volume as percentage of total email traffic from 2011 to 2021

Source: https://www.statista.com/statistics/420400/spam-email-traffic-share-annual/

Anti-spam filters seem to be the most viable solution nowadays. Most commercially available filters of this type currently appear to rely on simple techniques such as whitelists of trusted senders, blacklists of known spammers, and hand-crafted rules that block messages containing specific words or phrases (Ahmed et al, 2022). On the other hand, the success of Machine Learning in text categorisation has led researchers to explore learning algorithms in anti-spam filtering (Ahmed et al, 2022). Learning-based anti-spam filters, initially based on Naïve Bayes (NB), have been in operation for a few years (Michelakis et al, 2004). Recently, various other machine learning methods have been tested by researchers and demonstrate superior performance, as we shall see below.

In this paper, an algorithm for the detection of spam and phishing messages with the help of machine learning is proposed. The algorithm accepts as input text email messages written in English, grouped as benevolent ("ham") and malevolent (spam), and produces a set of training data in csv format (comma separated values) with the following columns: id, sender, subject, content and label. The fifth column is a binary variable characterising the messages. Next, the algorithm uses ten supervised learning methods that produce input-output mapping functions. Once trained, the algorithm can automatically recognise spam emails with high accuracy. In addition, the algorithm produces as output a file with the spam email sender Ipext, addresses, a csv file with statistics about the countries of origin, as well as a map with the geolocation of originating spam email servers. These files can be used to update existing blacklists and websites such as "Abuse IP DB" (https://www.abuseipdb.com), a project dedicated to helping combat the spread of hackers, spammers, and abusive activity on the internet.

The remainder of this paper is organised as follows: Section 2 reviews the related work on available anti-spam filters, focusing on learning-based and novel alternative approaches. Section 3 presents the datasets used. Section 4 presents the proposed system. Section 5 summarises the results of our algorithm on the benchmark datasets. Section 6 concludes the paper and suggests directions of further research.

2. Related work

Since identification is a specific task of classification, machine learning is usually used to classify email into two classes, namely spam and benign messages, also called "ham". Several studies using machine learning-based spam classification have been published in the research bibliography (Ghosh & Senthilrajan, 2023; Karyawati et al, 2023). Previous research work on anti-spam filtering studied the performance of popular machine learning algorithms (Michelakis et al, 2004; Ahmed et al, 2022; Ghosh & Senthilrajan, 2023).

Machine learning is easy to implement and good for high-dimensional data. However, it requires large and unbiased amounts of data. In addition, parameter adjustment is required to select the best model (Karyawati et al, 2023).

Laorden (2012) developed a Word Sense Disambiguation preprocessing step before applying machine learning algorithms to detect spam data. Results indicated a 2 to 6% increase in the precision score when applied on Ling Spam and TREC datasets.

Several studies have used Support Vector Machines (SVM) for spam classification. Jáñez-Martino et al (2020) made a combined model of TF-IDF (Term Frequency - Inverse Document Frequency) and SVM showed a 95.39% F1 score and the fastest spam classification was achieved with the help of the TF-IDF and NB approach (Roiger, 2017; Tan et al, 2006).

Deep learning is a subfield of machine learning that uses neural networks to learn patterns in data, mimicking the human brain. Marie-Saint et al (2020) employed deep learning to solve the given firefly algorithm with SVM and Arabic text.

Nonlinear kernel functions are often used to convert the input data to a high-dimensional feature space where the input data becomes more separable than the original input space (Karyawati et al, 2023).

Pratiwi and Ulama (2016) compared k-Nearest Neighbour (kNN), linear kernel SVM and Radial Basis Function (RBF) kernel SVM methods. It was found that the kNN method with k=3 produced the best accuracy of 92.28% while the best accuracy in the SVM method (96.6%) was obtained using the SVM linear kernel.

Chrismanto and Lukito (2017) compared the Naïve Bayes method and the SVM method with the RBF kernel to identify Instagram comment spam. The results showed that the SVM method produced an accuracy of 78.49%, outperforming Naïve Bayes which achieved an accuracy of 77.25%.

Pratiwi and Ulama (2016) proposed a combination of kNN and SVM methods. They used a kNN-based sampling strategy to find close neighbors to improve the performance of the SVM method. The results of the study based on publicly available dataset (Dredze) showed that the accuracy increased to about 98%.

Reddy and Reddy (2021) proposed a new spam detection method that is effective in distinguishing spam from its content. The proposed classifier achieved an accuracy of 95.32%.

Spammers have introduced a new technique of embedding spam content into images attached to email messages. Harisinghaney et al (2014) implemented text and image-based spam emails with the help of the kNN, NB and reverse DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithms. Similarly, Zamil et al (2019) proposed a method for distinguishing the ham and spam images using a combination of SVM and kNN, achieving an average accuracy of 97.27% when K equals 20.

Hemalatha et al (2022) discussed the use of the Naïve Bayes classifier to detect spam. Their classifier used techniques like lemmatisation, word removal and lowercasing, to analyse the text of the email in order to determine whether it is spam or not. The RFD model has also been used in anti-spam filters. This model can identify spam patterns, non-spam patterns and common patterns to determine whether an email is spam or ham. The RFD model offers general patterns that users can use to decide whether an email is spam or not. The proposed system is currently for English language emails but could be expanded to other languages in the future. Additionally, the study discussed how images can be used in spam emails and how characteristics like file properties, histogram, and hough change can be used to detect them. The text emphasized the importance of detecting spam to avoid loss of important mail but also protect user systems from harm.

Benavides et al (2020) reviewed various deep learning techniques used for classifying phishing attacks. The study discussed the limitations of traditional approaches (such as rule-based and signature-based methods) to phishing detection. Next, the authors highlighted the potential of deep learning techniques for detecting phishing attacks. The article reviewed several studies that have employed deep learning techniques, including artificial neural networks, convolutional neural networks and recurrent neural networks (RNN) for detecting phishing attacks. The authors suggested that deep learning techniques can provide accurate classification of phishing attacks and outperform traditional approaches. The study also noted the limitations of current research, such as the lack of standardization in data sets and the need for more comprehensive evaluations of deep learning-based phishing detection solutions.

Karyawati et al (2023) investigated several kernel functions (i.e., linear, polynomial and RBF) to obtain the best SVM model for classifying spam emails. Some experiments were conducted to determine the effect of parameter

changes for each kernel function. The SVM performance was assessed using the Precision, Recall, and F-Measure metrics. They used a publicly available email dataset posted on Kaggle, which is based on the Enron corpus (https://www.kaggle.com/datasets/venky73/spam-mails-dataset). The total number of 1100 messages consists of 550 spam emails and 550 benign emails, of which 1000 emails were used as training data (i.e., to select the best model of each kernel), and 100 emails are used for testing (i.e., to evaluate the best model). The feature data used was word frequency. The data was in the form of text in csv format.

Karyawati et al. (2023) concluded that the evaluation of the best model of each kernel model (among five models: linear, polynomial of degree 1, 2, 3 and RBF kernel) using the testing data of identification Spam email showed that the linear kernel and the kernel of degree 1 polynomial produced the best Precision, Recall and F-Measure performances of 0.99, 0.99 and 98% respectively, compared to the RBF kernel which produced Precision, Recall and F1 score of 0.95, 0.95 and 0.94, respectively. The worst model was a degree 3 polynomial kernel which produced Precision, Recall and F1 score of 0.85, 0.79 and 0.78, respectively.

In a recent research, Ghosh and Senthilrajan (2023) implemented and compared thirteen state-of-the-art machine learning classifiers: Adaptive Booster, Artificial Neural Network, Bootstrap Aggregating, Decision Table, Decision Tree, J48, K-Nearest Neighbor, Linear Regression, Logistic Regression, Naïve Bayes, Random Forest, Sequential Minimal Optimization and SVM. They concluded that in terms of accuracy, the Random Forest classifier performs best and the performance of the Naïve Bayes classifier is substandard compared to the rest of the classifiers. The Random Forest (RF) classifier achieved the accuracy of 99.91% for the Spam Corpus and 99.93% for the Spambase datasets. The Naïve Bayes classifier achieved the accuracy of 87.63% for the Spam Corpus and 79.53% for the Spambase datasets.

Finally, Sahmoud and Mikki (2022) investigated the application of a powerful natural language processing (NLP) model that has been shown to be effective in a variety of NLP tasks, called BERT (Bidirectional Encoder Representations from Transformers) in detecting spam messages. The authors trained their model on many corpuses such as SpamAssassin corpus, SMS Spam Collection v.1, Enron corpus and the Ling-Spam corpus, and analysed the results. They reported very good results of using BERT in every corpus they tested (from 97.83% to 99.28%), outperforming other machine learning algorithms in many tasks. The authors attributed the high performance of their model to the fact that the algorithm can better understand the message context and correlate the words (Sahmoud & Mikki, 2022).

3. Training the model

Common datasets which have been used in the literature include: Enron Corpus, SpamAssassin Corpus, SMS Spam Detection Corpus, Ling-Spam Corpus (Sahmoud & Mikki, 2022). In our research, the following datasets were used to train the model:

1) A dataset from SpamAssasin containing 2551 ham and 501 spam emails available online from:

https://www.kaggle.com/datasets/veleon/ham-and-spam-dataset

2) The Enron1 dataset available online from: https://www2.aueb.gr/users/ion/data/enron-spam, containing 3228 ham and 3219 spam emails.

3) We also used a ready csv file available online from:

https://www.kaggle.com/code/mfaisalqureshi/email-spam-detection-98-accuracy/data

This csv file (spam.csv, 480.13 kB) prepared by Faisal Qureshi, contains 5157 unique instances of ham (87%) and spam (13%) messages.

4. The proposed system

The proposed system accepts two sets of emails as input (ham and spam), written in English. It uses preprocessing to prepare a csv file with selected fields of the messages, as well as their characterisation. Then it uses this csv file to feed a Machine Learning engine running ten different algorithms mentioned in the Abstract, which after training can be used to filter the user's incoming messages. Alternatively, our code also accepts ready csv files. As an output, the proposed algorithm generates a csv file containing the spam emails source IP addresses and the countries of origin, statistics in textual and graphical format (histogram), as well as a map with the geolocation of the spammers. The algorithm is written in Python. Figure 2 presents a block diagram of the proposed system.

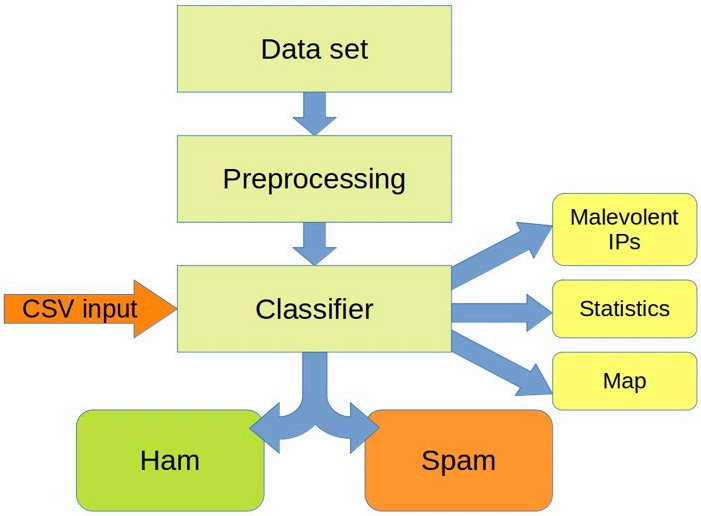


Figure 2: Block diagram of the proposed system

The code imports several Python libraries such as Json, Folium, Seaborn, Pandas and Nltk. It defines a function called "clean_text" which preprocesses email data by converting all the words to lowercase, removing URLs, punctuation marks, HTML tags, numbers and lemmatizing words. It also removes stopwords and filters out words that are less than four letters long. The function returns a clean text.

A function named "find_elements" extracts sender, subject, content and country information from emails and writes the extracted data into a csv file. Next, the "find_elements" function utilises regular expressions to extract IP addresses from emails and uses geolocation-db.com to obtain country names from IP addresses. Then, it calls the "clean_text" function to preprocess the email content. The function "create_stats_and_map" reads the csv file generated by the "find_elements" function and creates a map using the Folium library; it also uses a GeoJSON file that contains country boundaries and highlights each country according to the number of emails originating from it.

Figure 3 presents part of an output csv file. Figure 4 presents an example of the statistics, and Figure 5 presents a geolocation map with the places of the spammers of the Enron1 dataset.

	Sender	IP addr	Country
2	12a1mailbot1@web.de	203.122.2.197	India
3	ilug-admin@linux.ie	67.104.83.251	United States
4	sabrina@mx3.1premio.com	209.63.151.251	United States
5	wsup@playful.com	80.71.66.8	Denmark
6	social-admin@linux.ie	67.104.83.251	United States
7	Thecashsystem@firemail.de	24.214.168.210	United States
8	fort@bluemail.dk	217.127.249.196	Spain
9	OWNER-NOLIST-SGODAILY*JM**NETNOTE	209.216.124.212	Canada
10	Thecashsystem@firemail.de	24.214.168.210	United States
11	suz0123893616943@yahoo.com	65.113.29.188	United States
12	hurst@missouri.co.jp	202.101.163.34	China
13	simply-amateur-zzzz=example.com@free4por	205.252.89.51	United States
14	aileen@email2.qves.net	67.104.83.251	United States
15	OWNER-NOLIST-SGODAILY*JM**NETNOTE	168.143.181.42	United States
16	approvals@mindspring.com	200.173.221.83	Brazil
17	weseloh@bibsam.kb.se	200.24.85.9	Brazil
18	des34newsa@hotmail.com	63.186.16.221	United States
19	jjj@mymail.dk	61.97.34.233	South Korea
20	seko_mam@spinfinder.com	64.86.155.179	United States
21	safety33o@l11.newnamedns.com	64.25.38.81	United States
22	ilug-admin@linux.ie	209.63.151.251	United States
23	ilug-admin@linux.ie	64.86.155.148	United States
24	bell1hmed@yahoo.ca	64.86.155.148	United States
25	health104580m43@mail.com	135.5.185.20	United States
26	iq@insurancemail.net	65.217.159.103	United States
27	george300@Flashmail.com	61.129.68.17	China

Figure 3: Part of an output csv file

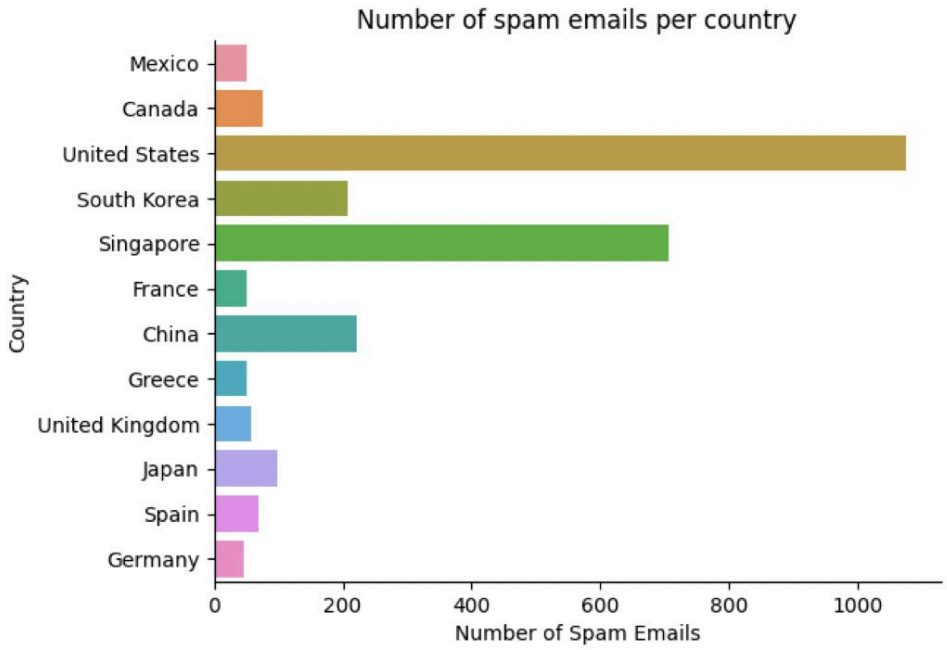


Figure 4: Statistics resulting from the spam emails (spam emails per country)

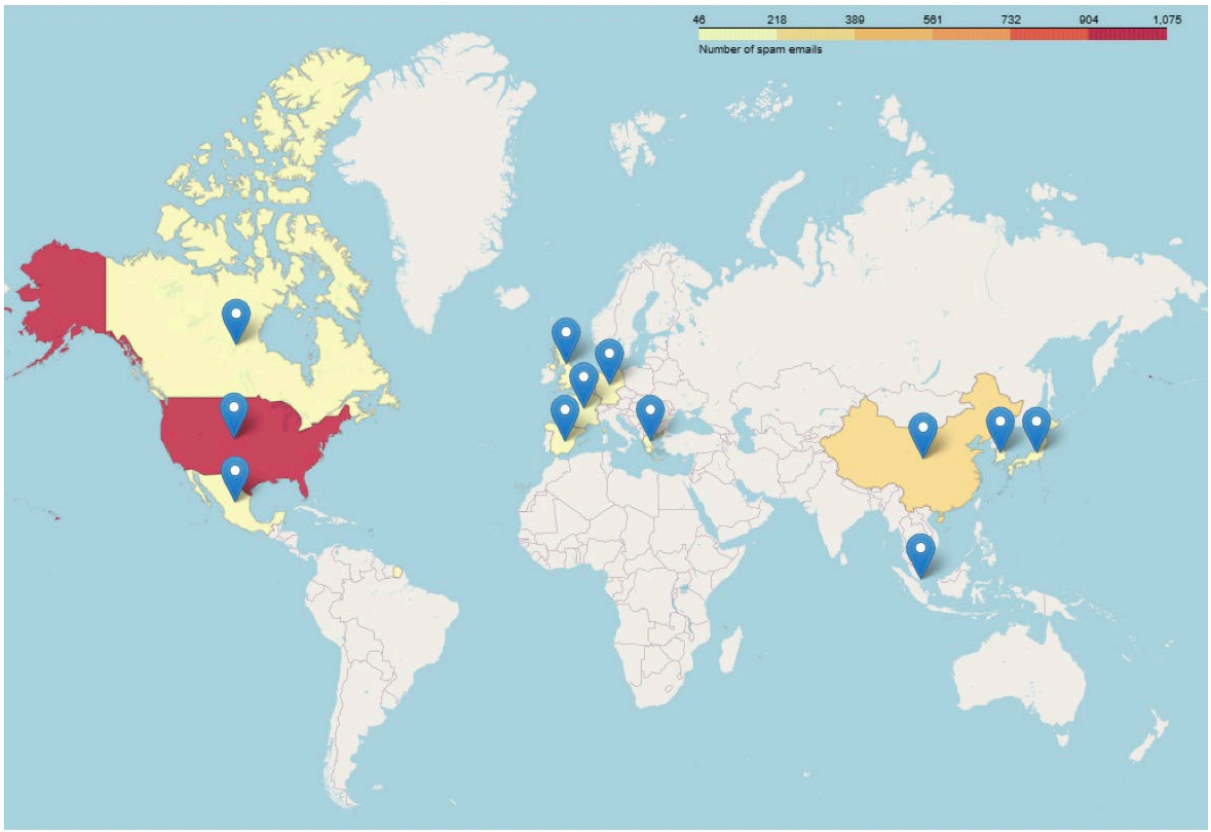


Figure 5: A geolocation map with the places of the spammers

5. Results

Table 1 shows our test results using several algorithms on the datasets under consideration, namely: Support Vector Machines (SVM), k-Nearest Neighbour (kNN, with k=1), Naïve Bayes (NB), Neural Networks (NN), Recurrent Neural Networks (RNN), Ada Boost (AB), Random Forest (RF), Gradient Boosting (RF), Logistic

Regression (LR) and Decision Trees (DT). The results are comparable and in some cases better than those reported in the recent bibliography (Ghosh & Senthilrajan, 2023; Ahmed et al, 2022).

Table 1: Accuracy of the proposed algorithm on various datasets under consideration

Method ► ▼ Dataset	SVM	kNN	NB	NN	RNN	AB	RF	GB	LR	DT
Spam Assasin	99.18%	95.9%	91.31%	99.51%	98.36%	98.2%	98.2%	95.9%	95.41%	95.73%
Enron1	99.38%	98.53%	98.3%	99.22%	99.3%	98.3%	97.2%	96.6%	99.22%	96.5%
csv by F. Qureshi	98.39%	95.51%	95.43%	98.23%	98.3%	98.38%	98.11%	95.9%	96.5%	97.85%

6. Conclusion

In this research work, an algorithm for the classification of raw emails consisting of spam and benign messages in English was presented. At a first stage, preprocessing produces a csv file with the basic email characteristics, which is next used to train ten popular machine learning classifiers to detect spam emails (Support Vector Machines, k-Nearest Neighbour, Naïve Bayes, Neural Networks, Recurrent Neural Networks, Ada Boost, Random Forest, Gradient Boosting, Logistic Regression and Decision Trees). Almost all classifiers demonstrate satisfactory performance, comparable to or superior than state-of-the-art implementations. Two publicly available datasets were used: the SpamAssasin dataset and the Enron1 dataset. Concerning the SpamAssasin dataset, the best performance (99.51 %) is achieved by NN. For the Enron1 dataset, the best performance (99.38 %) is achieved by SVM. Our algorithm also accepts ready csv files produced by other sources. We also tested our algorithm with a publicly available csv file; SVM and Ada Boost achieved the best performance (almost 98.4 %).

The proposed algorithm, implemented in Python, contains about 490 lines of code and can classify spam and ham emails. In addition, the proposed algorithm produces a csv file with the source IP addresses of the spam emails, statistics and graphs about the origin country of the spam emails, as well as a map with their geolocation. The IP addresses of the spam email senders can be used in spam email filters or to update spam email databases. In the future we intend to produce and test our own dataset, as well as work on spam filter personalisation.

References

Ahmed, N., Amin, R., Aldabbas, H., Koundal, D., Alouffi B. and Shah, T. (2022) "Machine Learning Techniques for Spam Detection in Email and IoT Platforms: Analysis and Research Challenges", *Hindawi Security and Communication Networks*, Vol. 2022, Article ID 1862888, https://doi.org/10.1155/2022/1862888

Benavides, E., Fuertes, W., Sanchez, S. and Sanchez, M. (2020) "Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review". In Á. Rocha and R. P. Pereira (eds.), *Developments and Advances in Defense and Security, Smart Innovation, Systems and Technologies 152*, Springer Nature Singapore Pte Ltd. https://doi.org/10.1007/978-981-13-9155-2_5

Chrismanto, A. R. and Lukito, Y. (2017) "Identifikasi Komentar Spam Pada Instagram", Lontar Komputer : Jurnal Ilmiah Teknologi Informasi, Vol. 8, no. 3, p. 219, doi: 10.24843/lkjiti.2017.v08.i03.p08.

Ghosh, A. and Senthilrajan, A. (2023) "Comparison of machine learning techniques for spam detection", *Multimedia Tools and Applications*, 1-28. 10.1007/s11042-023-14689-3.

Harisinghaney, A., Dixit, A., Gupta, S. and Arora, A. (2014) "Text and image based spam email classification using kNN, Naïve Bayes and Reverse DBSCAN algorithm", 2014 International Conference on Reliability Optimization and Information Technology (ICROIT) pp 153-155.

Hemalatha, M., Katta, S., Santosh, R. S. and Priyanka, Ms. (2022) "E-Mail Spam Detection", *International Journal of Computer Science and Mobile Computing*, Vol. 11, Issue 1, Jan. 2022, pp. 36-44.

Jáñez-Martino, F., Fidalgo, E., González-Martínez, S. and Velasco-Mata, J. (2020) "Classification of Spam Emails through Hierarchical Clustering and Supervised Learning," *CoRR*, Vol. abs/2005.08773.

Karyawati, A. E., Wijaya, K. and Supriana, I W. S. (2023) "A Comparison of Different Kernel Functions of SVM Classification Method for Spam Detection", *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, 8, pp 91-97, doi: 10.33480/jitk.v8i2.2463.

Laorden, C., Santos, I., Sanz, B., Alvarez, G. and Bringas, P. G. (2012) "Word Sense Disambiguation for Spam Filtering", *Electron. Commer. Rec. Appl.*, Vol. 11, pp 290–298.

Marie-Sainte, S. L. and Alalyani, N. (2020) "Firefly Algorithm based Feature Selection for Arabic Text Classification", *J. King Saud Univ. Comput. Inf. Sci.*, Vol. 32, pp 320-328.

Michelakis, E., Androutsopoulos, I., Paliouras, G., Sakkis, G., and Stamatopoulos, P. (2004) “Filtron: A Learning-Based Anti Spam Filter”, *International Conference on Email and Anti-Spam*.

Pratiwi, S. N. D. and Ulama, B. S. S. (2016) “Klasifikasi Email Spam dengan Menggunakan Metode Support Vector Machine dan k-Nearest Neighbor”, *Jurnal Sains dan Seni ITS*, Vol. 5, No. 2, pp 344–349.

Reddy, G. A. and Reddy, B. I. (2021) “Classification of Spam Text using SVM”, *Journal of University of Shanghai for Science and Technology*, Vol. 23, No. 08, pp 616–624, doi: 10.51201/jusst/21/08437.

Roiger, R. J. (2017) *Data Mining: A Tutorial-Based Primer*, 2nd ed. Boca Raton, CRC Press, Taylor & Francis Group.

Sahmoud, T. and Mikki, M. A. (2022) “Spam Detection Using BERT”, preprint 10.48550/arXiv.2206.02443.

SpamLaws.com (2023) [online], https://www.spamlaws.com [Accessed 6 March 2023].

Tan, P.-N., Steinbach, M. and Kumar, V. (2006) *Introduction to Data Mining*. Boston: Pearson Addison-Wesley.

Zamil, Y. K., Ali, S. A. and Naser, M. A. (2019) “Spam image email filtering using K-NN and SVM”, *International Journal of Electrical and Computer Engineering* (IJECE), Vol. 9, no. 1, pp 245-254, doi: 10.11591/ijece.v9i1.

Cyber Warfare and Cyber Terrorism Threats Targeting Critical Infrastructure: A HCPS-based Threat Modelling Intelligence Framework

Rennie Naidoo and Carla Jacobs
University of Pretoria, Pretoria, South Africa

rennie.naidoo@up.ac.za
u17053422@tuks.co.za

Abstract: Acts of cyber warfare and cyber terrorism (CWCT) that target a nation's critical infrastructure (CI) are quickly becoming a larger threat to national security than conventional kinetic warfare strategies. Adversaries or potential adversaries can target a nation's electrical grids, telecommunications, financial services, transportation, healthcare systems, and other forms of CI. These acts pose a major threat to a nation's CI and consequently exposes citizens to public health, safety, security, and economic development risks. Identifying cyber vulnerabilities and threats can help nations to improve their CI defence strategies. There is a crucial need for research that can aid in understanding the major types of CI threats and by what method they might occur. This paper conducts a systematic literature review to develop an initial threat intelligence framework of CWCT attacks on CI. Drawing from a Human–Cyber–Physical Systems (HCPS) lens, the threat intelligence framework classifies CWCT attacks according to the methods, weapons, vulnerabilities, targets and impact of the CWCT attack. The cyber warfare community can extend the proposed HCPS-based threat intelligence framework to develop more advanced cyber security mitigation strategies, training scenarios and simulations. Large-scale monitoring of CI threats requires in-depth threat intelligence analysis and a collaborative defence strategy. This calls for a higher degree of coordination and orchestration between the military, intelligence agencies, government departments, multinational allies, regulators, and commercial entities. Future research can customize the proposed HCPS-based threat intelligence framework to cater for the unique threats facing specific CI domains.

Keywords: Cyberwar; Cyber Terrorism; Human–Cyber–Physical Systems; Threat Modelling; Critical Infrastructure; Systematic Literature Review

1. Introduction

Cyberwarfare has become a major concern for many nations and is considered a significantly higher threat to national security than conventional kinetic warfare strategies (Kaiser, 2014). A recent global survey found that cyberattacks targeting critical infrastructure made up 40% of all nation-state attacks (Microsoft's Digital Defense Report, 2022). While a substantial portion of this share was made up of Russian state-sponsored cyberattacks targeting Ukrainian infrastructure in the ongoing Russian-Ukrainian war, some scholars argue that any cyberattack aimed at destroying CI systems is tantamount to a declaration of war (Indrajit et al, 2021). Cyberattacks poses a major threat to a country's electrical grid, telecommunications, financial services, transportation and healthcare systems. The impact of these attacks often extend beyond the targeted CI with the potential of causing significant collateral damage (Abouzakhar et al, 2018).

Many nations are now identifying threats and mitigation strategies that will enable them to better protect their CI from the attacks of adversaries or potential adversaries in cyberspace. Recently, President Biden enacted the Cyber Incident Reporting for Critical Infrastructure Act (CIRCIA, 2022) which reinforces the need to improve the threat modelling and intelligence capability of US-based organizations and EU allies. Given the increasing volume and diversity of threats targeting CI, spotting trends and quickly sharing intelligence is crucial for hardening and defending CI. A nation's threat modelling and intelligence capability can improve the safeguarding of CI by developing and strengthening their cyber capabilities and tactics.

According to Shi et al (2021), threat modelling is a systematic process for identifying, evaluating and developing countermeasures to protect critical assets from threats and threat actors. A crucial extension of threat modelling is cyber threat intelligence (Kotsias et al, 2022; Tounsi and Rais, 2018; Shackleford, 2017). Cyber threat intelligence refers to the process of "acquiring, processing, analysing, and disseminating information that identifies, tracks, and predicts threats, risks, and opportunities inside the cyber domain to offer courses of action that enhance decision making" (Ettinger, 2019). A number of threat models exist, such as STRIDE, DREAD, P.A.S.T.A, Trike, VAST, Attack Tree, Common Vulnerability Scoring System (CVSS), T-MAP and OCTAVE, to assist in identifying vulnerabilities and mitigating potential threats facing networks, computers, software products, and data (Shostack, 2014, Kotsias et al, 2022; Shi et al, 2021). However, these models for the most part focus at an organizational level. More recently, the Cyber-Physical Systems (CPSs) approach has been proposed for threat

modelling and analysis of specific CI domains (Ding et al, 2017; Lee 2015). However, these approaches tend to emphasise computational and the physical components without explicitly integrating the human component into their models which may be too narrow for the holistic analysis of CWCT attacks targeting national CI domains (Xiong and Lagerström, 2019).

This research paper draws from recent research on Human–Cyber–Physical Systems (HCPSs) to provide an initial threat intelligence framework that identifies the different types of cyber warfare and cyber terrorism (CWCT) attacks on CI across the globe (Zhou et al, 2019). The research surveys recent literature to analyse the different types of attacks that have been observed in recent years with the aim of classifying them according to distinct factors such as method, weapons, vulnerabilities, and targets of the CWCT attack. The study also briefly discusses some of the methods and tactics that can be adopted by cybersecurity practitioners to harden CI.

The rest of the paper is organized as follows: first, we outline the HCPS-based framework as a basis for our analysis. Second, we present our systematic literature approach to review the selected CI literature in more detail. We then present and discuss the results and finally conclude the paper.

2. Conceptual Foundations

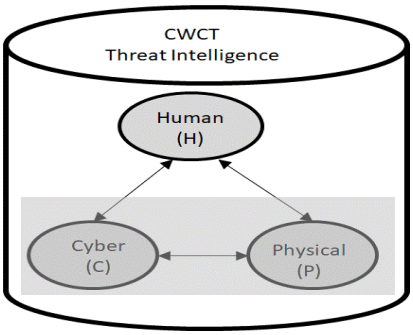


Figure 1: HCPS-based threat intelligence modelling framework

A popular way to think about CI cybersecurity is to view it as a cyber-physical system (CPS). While CPS is sometimes conflated with the term cybersecurity, CPS is more encompassing as it entails all interactions between the cyber and physical environment. The architecture of a cyber-physical systems (CPS) includes digital, analogue and physical components (Lee 2015). From an intellectual standpoint, the dominant approaches to study CPS is multidisciplinary but tends to be limited to the computer science and the engineering disciplines. CPS has been valuable to the study of the cybersecurity of CIs as it considers the dynamic interaction among computers, networking, and physical systems domains (Ding et al, 2017). However our study is aligned to more recent advances in CPS studies that are now explicitly featuring humans and human systems in the so-called Human–Cyber–Physical Systems (HCPSs) (Zhou et al, 2019). We propose a HCPS-based approach to frame threat modelling and intelligence of CWCT that target a nation's CI. To cater for the CI environment, we employ the term threat intelligence more broadly as the use of sensitising concepts to aid in thinking, learning and disseminating information about cybersecurity threats in HCPSs. Given the inherent complexity of HCPS landscape, we limit our scope to an asset and attacker centric approach to threat intelligence modelling (Tounsi and Rais, 2018; Shostack, 2014). Specifically, we pay attention to the attack target, cyber weapon or method used in the attack, vulnerabilities, and impact of the attack in a HCPS.

3. Research Method

This study has been conducted using a systematic literature review research method. Articles that focused on cyberwar or cyber-attacks on CI were included. The following search string (Cyberwar OR "Cyber War") AND (Attack OR Threat) AND "Critical Infrastructure") OR (Cybersecurity OR “Cyber Security” OR “Cyber-Attack”) AND “Critical Infrastructure”). The following databases were searched using the defined search string: IEEE Xplore® Digital Library, JSTOR and Scopus. A filtering process was used to ensure that the academic journal articles included in the systematic literature review are of high quality and relevant to the research question. Articles were first identified by using the search string. Thereafter, the duplicate articles were removed, and the abstracts of the remaining articles were screened. After the abstract screening step, the full text for each article was then assessed for eligibility. The following table displays the results after each of the filtration steps as described.

Table 1: Results of the Filtering Process

Filter process steps	Results
Identify articles	405
Remove duplicates	391
Screen abstract	391
Screen full text	72
Studies included	44

During the data extraction process, the following dimensions have been extracted from each of the data sources: CWCT attack target, cyber weapon or method used in the attack, vulnerabilities, and impact of the attack. The purpose of extracting each of the abovementioned dimensions from each of the included research papers was to analyse and categorise the findings. The research papers included in the study were analysed according to which CI component was targeted by a cyber-attack, the vulnerabilities of a cyber-attack and the potential impact on humans, the environment, and the economy. The texts for each of the data sources included in the literature review were analysed to extract meaningful relationships that would aid in answering the research question. Certain limiting factors, such as secrecy and the ever-evolving nature of cyber risks, may affect the accuracy of the research findings.

4. Results

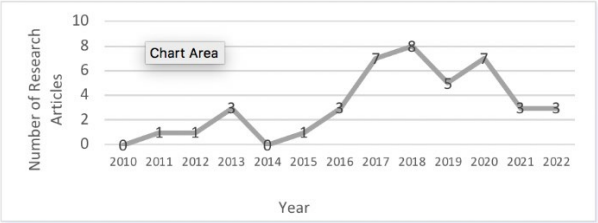


Figure 2: Cybersecurity Research Trends

Figure 2 depicts the upward trend in cybersecurity research focusing on CI components completed in the past twelve years. Although the graph is limited to the research articles included in this study, there was notable growth in cyberwar research from the year 2014 to the year 2018. To identify the components in CI that are most vulnerable to CWCT attacks, the following bar graph visually depicts in how many of the research papers each type of identified cyber-attack target was mentioned. It is important to note that some attack targets are components or systems within CIs, such as Supervisory Control and Data Acquisition (SCADA) systems, which monitors and controls Smart Grids.

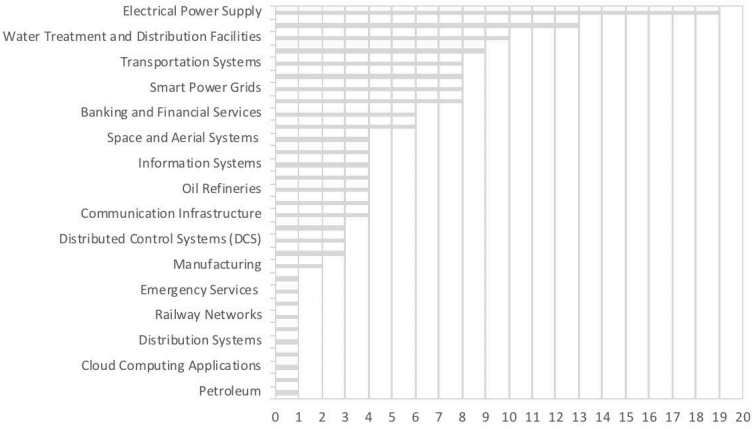


Figure 3: Critical Infrastructure CWCT Attack Targets

This literature review is directed towards identifying which types of CWCT attacks pose the biggest threat to CI. Therefore, the subsequent analysis of this research paper was completed using only the cyber-attack targets

that were mentioned in most research articles. Based on the findings depicted in the above graph, the CI systems and components of CI systems that are most vulnerable to cyber-attacks are the following: Banking and Financial Services, Cyber Physical Systems, Transportation Systems, Water Treatment and Distribution Systems, Nuclear Power Plants, Industrial Control Systems, Smart Power Grids, Supervisory Control and Data Acquisition (SCADA) Systems, Electrical Power Supply, and Hospitals and Healthcare. The identified cyber-attack weapons or methods that are used in most cyberwar attacks against the identified major targets are depicted in Figure 4.

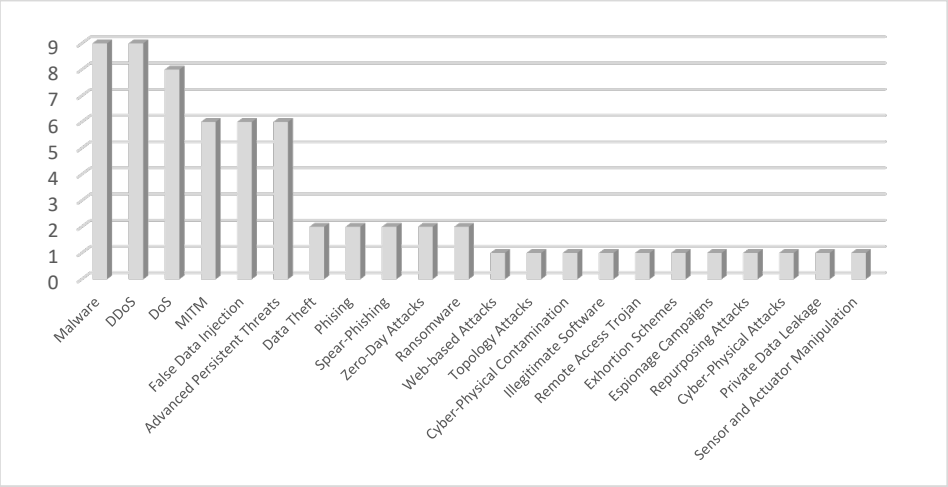


Figure 4: Cyber Weapons Mentions

The major types of cyberwar attack weapons or methods of attack are Malware, Denial of Service (DoS) and Distributed Denial of Service (DDoS) attacks, Man-in-the-Middle (MITM) attacks, False Data Injections (FDI) and Advanced Persistent Threats (APTs).

Figure 5 shows the identified cyber weapons used in CWCT attacks. Each attack target also highlights the impact that a potential CWCT attack could have on the normal operation of that CI component and the identified vulnerabilities which must be protected to safeguard CIs.

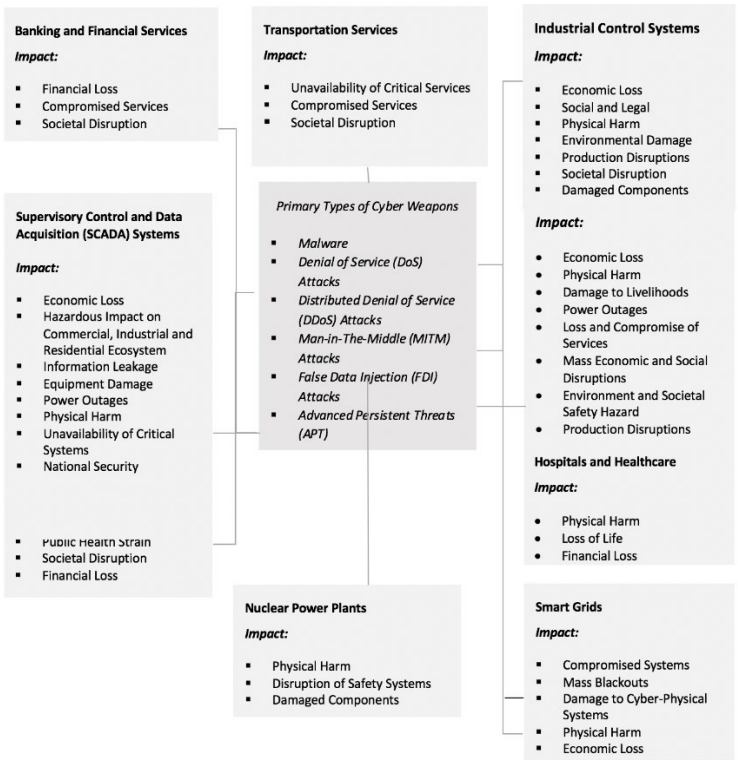


Figure 5: Impact of CWCT Cyber Threats on Critical Infrastructure

The following table contains tactics that can prevent CWCT attacks on CI from occurring or which can help to limit the amount of damage that might be caused from such an attack.

Table 2: Tactics to Address CWCT Attacks

Attack Prevention and Damage Control Tactics	References
Sensor Data Protection; Attack Detection Mechanisms; Intrusion Detection; Cryptographic Keys; Combining Cryptography and Steganography; Context-Aware Security Frameworks	*(Ashibani & Mahmoud, 2017)*
Remote Access Control Mechanisms; Data Encryption	*(Abouzakhar et al, 2018)*
Virtual Testbeds	*(Alves et al, 2016)*
Machine Learning; Artificial Intelligence	*(Nguyen et al, 2020)*
Anomaly Detection; Extended Firewall Usage; Data Leakage Protection	*(Domínguez et al, 2017)*

The main attack prevention and damage control tactics that were identified from the literature includes attack detection mechanisms, anomaly detection, data encryption, and use of cryptographic keys. Nguyen et al (2020) mentions that machine learning and artificial intelligence techniques can also be used to identify potential cyber-attacks. Multiple stakeholders including the military, intelligence agencies, government departments, multinational allies, regulators, and commercial entities will need to establish close partnerships develop the threat intelligence capability required to secure CI.

5. Discussion

The focus of this study has been on identifying the major types of cyberwar threats facing CI. Certain CI components and systems within them more important to national security than others, and greater attention must be given to ensuring the protection and functioning of these. CI such as the electrical power supply and smart power grids are the backbone of many other national CIs such as transportation systems, healthcare, and water distribution (Adepu et al, 2020). In the research findings, electrical power supply was identified as the most vulnerable CWCT attack target. This is a major challenge given that this CI is also the backbone on which many other CIs such as healthcare and banking services are entirely reliant.

The protection of cybersecurity systems in the healthcare and water distribution sector is crucial because attacks on these have serious public health consequences. Based on the findings in literature, humans can experience severe health risks or even loss of life when a country's healthcare infrastructure becomes the target of a major cyberwar attack. According to Kendzierskyj and Jahankhani (2019), there has been a significant increase in data breaches and ransomware attacks in the healthcare sector since 2016, the primary reason being that the systems are in critical need of cybersecurity enhancements. Delays or interruptions in these systems would have a major impact on healthcare operations. The major types of CWCT attacks against healthcare infrastructure that were identified from the data analysis were Denial of Service (DoS) attacks, Malware, Private Data Leakage and Data Theft. The consequences of these attacks are patients suffering physical harm and financial loss to the patients and hospitals. In some cases, attacks on hospital infrastructure were even classified as life-threatening (Abouzakhar et al, 2018). Some of the attack prevention tactics identified by Abouzakhar et al (2018) to impede Denial of Service attacks on hospitals were the implementation of effective remote access control mechanisms, proper security configuration and intrusion detection.

As identified in the findings, traditional electrical power grids and modern smart power grids are a major CI cybersecurity concern. Unlike traditional electrical power grids, communication networks play a crucial role in the smart grid environment where components such as sensors and controllers are continually communicating with and exchanging data between each other (Gupta & Akhtar, 2017). Due to the heavy interconnectivity of the components in smart power grids, they are more susceptible to cyber-attacks than the traditional manual power grids. Although their connectivity and intelligence make them an appealing target for potential CWCT attacks, it also minimises the impacts that an attack may have, because they can be reconfigured without human intervention due to their interconnected nature. This capability of the smart grid can play an important role in the prevention of negative cyber-attack outcomes, such as mass electrical outages (Gupta & Akhtar, 2017). Some extended security measures identified by Domínguez et al (2017) for the protection of smart grids and other CIs include anomaly detection, extended firewall usage, and data leakage protection.

Industrial control systems have also been identified as a major cyberwar attack target. To elaborate on one of the identified vulnerabilities in industrial control systems, Domínguez et al (2017) identified a mismatch between industrial control experts' knowledge in the field of computer and cyber security, and cybersecurity experts' knowledge of industrial control system operations. According to Domínguez et al (2017), industrial control experts generally lack computer security training and cybersecurity experts tend to disregard industrial control system operations, which leaves a concerning knowledge and skills gap that must be addressed by appropriate training environments (Domínguez et al, 2017). Another serious vulnerability can be seen for industrial control systems (ICS), which were originally designed for secluded operations. Because they were designed as isolated systems, their immersion into the world of interconnectivity exposes these systems to several alarming external cybersecurity threats that were not anticipated previously (Domínguez et al, 2017).

Supervisory Control and Data Acquisition (SCADA) systems were identified as a key component in CIs. According to Huang et al (2017), the protection of SCADA systems are a major national cybersecurity concern, as they are a crucial component of industrial applications like smart grids. SCADA systems are used in the monitoring and controlling of critical processes, and attacks against these could have numerous consequences on a country, such as power outages and information leakages. Samanti et al (2016) argues that SCADA Systems are primarily under security threats from insiders because SCADA systems and networks have largely remained isolated from the internet for most of their existence. A method to address the CWCT attacks on SCADA systems was identified by Nguyen et al (2020), which involves using machine learning and artificial intelligence techniques to identify potential cyber-attacks.

6. Conclusion

This study surveyed the major types of cyber warfare and cyber terrorism (CWCT) threats to CI by examining recent types of cyber threats that were aimed at the different types and components of national CI. This literature review built on previous research by developing an initial threat intelligence framework drawing from key features of a Human-Cyber–Physical-Systems (HCPS). This particular study identifies the different types of CWCT attacks on CI observed in recent years and classifies each of these attacks according to factors including method, weapons used, vulnerabilities, and targets of each of the CWCT attacks. Two of the CI components that were identified as having the highest risk of being targeted by CWCT attacks were the electrical power supply infrastructure and related SCADA system components. Tactics to address the threats facing CI include creating training environments to address cybersecurity knowledge gaps, implementing proper security configuration and intrusion detection methods, and using machine learning and artificial intelligence techniques to identify potential cyber-attacks. Threat mitigation also requires a high degree of coordination and orchestration between the military, intelligence agencies, government departments, multinational allies, regulators, and commercial entities. Researchers in the field of cyberwar research can expand on and customize the proposed CI CWCT threat intelligence framework to cater for the unique threats facing specific CI components. Our future work will extend the CWCT threat intelligence framework to cater for other critical elements of a HCPS.

References

Sources reviewed in the SLR

Abouzakhar, N. S., Jones, A., & Angelopoulou, O. (2018). Internet of Things Security: A Review of Risks and Threats to Healthcare Sector. *Proceedings - 2017 IEEE International Conference on Internet of Things, IEEE Green Computing and Communications, IEEE Cyber, Physical and Social Computing, IEEE Smart Data, iThings-GreenCom-CPSCom-SmartData 2017.* DOI: 10.1109/iThings-GreenCom-CPSCom-SmartData.2017.62

Adepu, S., Kandasamy, N. K., Zhou, J., & Mathur, A. (2020). Attacks on smart grid: power supply interruption and malicious power generation [Article]. *International Journal of Information Security*, *19*(2), 189-211.

Ahmad, R., & Yunos, Z. (2012). A dynamic cyber terrorism framework. *International Journal of Computer Science and Information Security, 10(2), 149.*

Akhtar, T., Gupta, B. B., & Yamaguchi, S. (2018, 2018). Malware propagation effects on SCADA system and smart power grid. *IEEE International Conference on Consumer Electronics (ICCE)* https://dx.doi.org/10.1109/icce.2018.8326281

Alves, T., Das, R., & Morris, T. (2016). Virtualization of industrial control system testbeds for cybersecurity. *ACM International Conference Proceeding Series*

Arnold, N. R., Mahoney, W. R., Derrick, D. C., Ligon, G. S., & Harms, M. M. (2015). Feasibility of a Cyber Attack on National Critical Infrastructure by a Non-State Violent Extremist Organization. *Journal of Information Warfare*, *14*(1), 84-100.

Ashibani, Y., & Mahmoud, Q. H. (2017). Cyber physical systems security: Analysis, challenges and solutions [Article]. *Computers and Security*, *68*, 81-97.

Ashrafuzzaman, M., Jamil, H., Chakhchoukh, Y., & Sheldon, F. (2018). A Best-Effort Damage Mitigation Model for Cyber-Attacks on Smart Grids. *IEEE International Conference on Computer Software & Applications.* https://dx.doi.org/10.1109/compsac.2018.10285

Axelrod, C. W. (2013). Managing the risks of cyber-physical systems. 9th Annual Conference on Long Island Systems, Applications and Technology, LISAT 2013, DOI: 10.1109/LISAT.2013.6578215

Barmin, Y., Jones, G., Moiseeva, S., & Winkelman, Z. (2011). International Arms Control and Law Enforcement in the Information Revolution An Examination of Cyber Warfare and Information Security. *Connections*, *10*(4), 73-94.

Benameur, A., Evans, N. S., & Elder, M. C. (2013). Cloud resiliency and security via diversified replica execution and monitoring. *Proceedings - 2013 6th International Symposium on Resilient Control Systems, ISRCS 2013.* DOI: 10.1109/ISRCS.2013.6623768

Bhamare, D., Zolanvari, M., Erbad, A., Jain, R., Khan, K., & Meskin, N. (2020). Cybersecurity for industrial control systems: A survey [Review]. *Computers and Security*, *89*, Article 101677.

Botes, M., & Lenzini, G. (2022). When Cryptographic Ransomware Poses Cyber Threats: Ethical Challenges and Proposed Safeguards for Cybersecurity Researchers. *Proceedings - 7th IEEE European Symposium on Security and Privacy Workshops, Euro S and PW 2022*, 562-568. DOI: 10.1109/EuroSPW55150.2022.00067

Branquinho, M. A. (2018). Ransomware in industrial control systems. What comes after Wannacry and Petya global attacks? *WIT Transactions on the Built Environment, 329-334*. DOI: 10.2495/SAFE170301

Busquim e Silva, R. A., Piqueira, J. R. C., Cruz, J. J., & Marques, R. P. (2021). Cybersecurity Assessment Framework for Digital Interface Between Safety and Security at Nuclear Power Plants [Article]. *International Journal of Critical Infrastructure Protection*, *34*, Article 100453.

Chen, Z., Yan, L., He, Y., Bai, D., Liu, X., & Li, L. (2018). Reflections on the Construction of Cyber Security Range in Power Information System. *IEEE.* https://dx.doi.org/10.1109/iaeac.2018.8577685

Djenna, A., Saidouni, D. E., & Abada, W. (2020). A pragmatic cybersecurity strategies for combating IoT-cyberattacks. *International Symposium on Networks, Computers and Communications, ISNCC 2020*. DOI: 10.1109/ISNCC49221.2020.9297251

Domínguez, M., Prada, M. A., Reguera, P., Fuertes, J. J., Alonso, S., & Morán, A. (2017). Cybersecurity training in control systems using real equipment. DOI: 10.1016/j.ifacol.2017.08.2151

Eder-Neuhauser, P., Zseby, T., Fabini, J., & Vormayr, G. (2017). Cyber attack models for smart grid environments [Article]. *Sustainable Energy, Grids and Networks*, *12*, 10-29.

Gupta, B. B., & Akhtar, T. (2017). A survey on smart power grid: frameworks, tools, security issues, and solutions [Article]. *Annales des Telecommunications/Annals of Telecommunications*, *72*(9-10), 517-549.

Hathaway, O. A., Crootof, R., Levitz, P., Nix, H., Nowlan, A., Perdue, W., & Spiegel, J. (2012). The Law of Cyber-Attack. *California Law Review*, *100*(4), 817-885.

Huang, K., Zhou, C., Tian, Y.-C., Tu, W., & Peng, Y. (2017, 2017). Application of Bayesian network to data-driven cyber-security risk assessment in SCADA networks. *IEEE.* https://dx.doi.org/10.1109/atnac.2017.8215355

Huseinovic, A., Mrdovic, S., Bicakci, K., & Uludag, S. (2018). A Taxonomy of the Emerging Denial-of-Service Attacks in the Smart Grid and Countermeasures. *2018 26th Telecommunications Forum, TELFOR 2018 - Proceedings.* DOI: 10.1109/TELFOR.2018.8611847

Indrajit, R. E., Marsetio, Gultom, R., & Widodo, P. (2021). Cyber Troops: Developing Collective Abilities to Face Cyberwarfare Challenges. *IEEE.* https://dx.doi.org/10.1109/icacsis53237.2021.9631306

Jiang, Y., Jeusfeld, M., Atif, Y., Ding, J., Brax, C., & Nero, E. (2018). A Language and Repository for Cyber Security of Smart Grids. *IEEEI.* https://dx.doi.org/10.1109/edoc.2018.00029

Kaiser, R. (2015). The birth of cyberwar. *Political Geography, 46, 11-20.* https://www.researchgate.net/publication/273113188_The_birth_of_cyberwar

Kant, D., Creutzburg, R., & Johannsen, A. (2020). Investigation of risks for critical infrastructures due to the exposure of SCADA systems and industrial controls on the Internet based on the search engine Shodan. *IS and T International Symposium on Electronic Imaging Science and Technology.* DOI: 10.2352/ISSN.2470-1173.2020.3.MOBMU-253

Kendzierskyj, S., & Jahankhani, H. (2019, 2019). The Role of Blockchain in Supporting Critical National Infrastructure. *IEEE.* https://dx.doi.org/10.1109/icgs3.2019.8688026

Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology*, *51*(1), 7-15.

Kobara, K. (2016). Cyber physical security for Industrial Control Systems and IoT [Article]. *IEICE Transactions on Information and Systems*, *E99D*(4), 787-795.

Kozik, R., & Choras, M. (2013). Current cyber security threats and challenges in critical infrastructures protection. *IEEE.* https://dx.doi.org/10.1109/icoia.2013.6650236

Kshetri, N., & Voas, J. (2017). Hacking Power Grids: A Current Problem [Article]. *Computer*, *50*(12), 91-95, Article 8220480.

Kure, H. I., Islam, S., & Razzaque, M. A. (2018). An integrated cyber security risk management approach for a cyber-physical system [Article]. *Applied Sciences (Switzerland)*, *8*(6), Article 898.

Laszka, A., Abbas, W., Vorobeychik, Y., & Koutsoukos, X. (2017). Synergic security for smart water networks: Redundancy, diversity, and hardening. *Proceedings - 2017 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks, CySWATER 2017.* DOI: 10.1145/3055366.3055376

Libicki, M. C. (2011). Cyberwar as a Confidence Game. *Strategic Studies Quarterly, 5(1), 132–147.* http://www.jstor.org/stable/26270514

Li, F., Yan, X., Xie, Y., Sang, Z., & Yuan, X. (2019, 2019). A Review of Cyber-Attack Methods in Cyber-Physical Power System. *IEEE.* https://dx.doi.org/10.1109/apap47170.2019.9225126

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med 6(7): e1000097*. doi:10.1371/journal.pmed1000097

Nguyen, T., Wang, S., Alhazmi, M., Nazemi, M., Estebsari, A., & Dehghanian, P. (2020). Electric Power Grid Resilience to Cyber Adversaries: State of the Art. *IEEE Access*, *8*, 87592-87608.

Oughton, E. J., Ralph, D., Pant, R., Leverett, E., Copic, J., Thacker, S., Dada, R., Ruffle, S., Tuveson, M., & Hall, J. W. (2019). Stochastic Counterfactual Risk Analysis for the Vulnerability Assessment of Cyber-Physical Attacks on Electricity Distribution Infrastructure Networks [Article]. *Risk Analysis*, *39*(9), 2012-2031.

Park, J. W., & Lee, S. J. (2019). Probabilistic safety assessment-based importance analysis of cyber-attacks on nuclear power plants [Article]. *Nuclear Engineering and Technology*, *51*(1), 138-145.

In Boland, A., In Cherry, M. G., & In Dickson, R. (2017). Doing a systematic review: A student's guide.

Ramotsoela, T. D., Hancke, G. P., & Abu-Mahfouz, A. M. (2020). Behavioural Intrusion Detection in Water Distribution Systems Using Neural Networks [Article]. *IEEE Access*, *8*, 190403-190416.

Samanis, E., Gardiner, J., & Rashid, A. (2022). Adaptive Cyber Security for Critical Infrastructure. *IEEE.* https://dx.doi.org/10.1109/iccps54341.2022.00043

Samtani, S., Yu, S., Zhu, H., Patton, M., & Chen, H. (2016). Identifying SCADA vulnerabilities using passive and active vulnerability assessment techniques. *IEEE International Conference on Intelligence and Security Informatics: Cybersecurity and Big Data, ISI 2016, 25-30.* DOI: 10.1109/ISI.2016.7745438

Semertzis, I., Rajkumar, V. S., Stefanov, A., Fransen, F., & Palensky, P. (2022, 2022). Quantitative Risk Assessment of Cyber Attacks on Cyber-Physical Systems using Attack Graphs. *IEEE.* https://dx.doi.org/10.1109/mscpes55116.2022.9770140

Seyam, A. R., Nassif, A. B., Nasir, Q., Al Blooshi, B., & Talib, M. A. (2021). Deep learning techniques to detect DoS attacks on industrial control systems: A systematic literature review. *ACM International Conference Proceeding Series 2021.* DOI: 10.1145/3485557.3485577

Shrestha, M., Johansen, C., Noll, J., & Roverso, D. (2020). A Methodology for Security Classification applied to Smart Grid Infrastructures [Article]. *International Journal of Critical Infrastructure Protection*, *28*, Article 100342.

Stoddart, K. (2016). UK cyber security and critical national infrastructure protection. *International Affairs (Royal Institute of International Affairs 1944-)*, *92*(5), 1079-1105.

Van Epps, G. (2013). Common Ground Connections U.S. and NATO Engagement with Russia in the Cyber Domain. *Connections*, *12*(4), 15-50.

Vessels, L., Heffner, K., & Johnson, D. (2019). Cybersecurity risk assessment for space systems. *Proceedings - 2019 IEEE Space Computing Conference, SCC 2019, 11-19*. DOI: 10.1109/SpaceComp.2019.00006

Ghafir, I., & Prenosil, V. (2014). Advanced persistent threat attack detection: an overview. *Int J Adv Comput Netw Secur, 4(4), 5054.*

Other references

CIRCIA (2022) “Cyber Incident Reporting for Critical Infrastructure Act of 2022”, [online], https://www.cisa.gov/sites/default/files/publications/Cyber-Incident-Reporting-ForCriticalInfrastructure-Act-of2022_508.pdf (accessed on 12 January 2023).

Ding, J., Atif, Y., Andler, S. F., Lindström, B., & Jeusfeld, M. (2017). CPS-based threat modeling for critical infrastructure protection. ACM SIGMETRICS Performance Evaluation Review, 45(2), 129-132.

Ettinger, J. (2019). Cyber intelligence tradecraft report: The state of cyber intelligence practices in the United States. Retrieved from Carnegie Mellon University: https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=546686 (accessed on 12 January 2023).

Kotsias, J., Ahmad, A., & Scheepers, R. (2022). Adopting and integrating cyber-threat intelligence in a commercial organisation. European Journal of Information Systems, 1-17.

Lee, E. A. (2015). The past, present and future of cyber-physical systems: A focus on models. Sensors, 15(3), 4837-4869.

Microsoft Digital Defense Report (2022). Microsoft, https://www.microsoft.com/en-us/security/business/microsoft-digital-defense-report?rtc=1. (accessed on 11 January 2023).

Shackleford, D. (2017). Cyber threat intelligence uses, successes and failures: The SANS 2017 CTI survey. SANS Institute.

Shi, Z., Graffi, K., Starobinski, D., & Matyunin, N. (2021). Threat Modeling Tools: A Taxonomy. IEEE Security & Privacy, (01), 2-13.

Shostack, A. (2014). Threat modeling: Designing for security. John Wiley & Sons.

Tounsi, W., & Rais, H. (2018). A survey on technical threat intelligence in the age of sophisticated cyber attacks. Computers & security, 72, 212-233.

Xiong, W., & Lagerström, R. (2019). Threat modeling–A systematic literature review. Computers & security, 84, 53-69.

Zhou, J., Zhou, Y., Wang, B., & Zang, J. (2019). Human–cyber–physical systems (HCPSs) in the context of new-generation intelligent manufacturing. Engineering, 5(4), 624-636.

Determination of the end Device risk Likelihood Using the Bayesian Network Tools

Tabisa Ncubukezi
Information Technology Department, Faculty of Informatics and Design
Cape Peninsula University of Technology, Cape Town, South Africa

Ncubukezit@cput.ac.za

Abstract: All institutions use end devices for information processing which includes sending and receiving on the network. This process helps them to improve their business production as well as perform daily activities at a faster rate. However, the increased usage of end devices by both employees and criminals raises concerns and exposes businesses to a range of cyber risks. End devices can sometimes be used as agents and weapons to expose internal business operations. The vulnerability of the end devices to cyber threats and attacks compromises business data, its safety, and security. This paper determines the risk likelihood of the end devices using the Bayesian network tools. To achieve this, the study illustrates the connections of the end device variables to simulate the risk likelihood and its impact. The analysis and interpretation of the simulation are performed using decision tree analysis (DTA), scenario analysis, and sensitivity analysis techniques (Tornado graphs, conditional probability tables (CPT), and value of information configuration (VOI)). The relationship of the variables is demonstrated on the AgenaRisk package. Results revealed variables that influence the risk probability and its impact. End device risks can be caused by insiders and cybercriminals. The risks associated with end devices are influenced by the level of security implementation on different levels. The impact of the cyber risks was also accounted for and the concluding remarks were also made.

Keywords: AgenaRisk, Bayesian Network, Cybersecurity, Decision Tree Analysis, End Devices, Risk Assessment, Risk Likelihood, Sensitivity Analysis

1. Introduction

For communication purposes, the network is made up of different levels which include devices. The devices can be positioned as the end nodes, interconnections, and peripherals that form part of the business systems. Devices connected to the Internet or acting as stand-alone may be the main entry points that victims use to commit cybercrimes. Cybercrimes can either be caused by intentional cybercriminals or ignorant employees (Ncubukezi, 2022a). All business institutions are operated by both computer-literate and illiterate employees and are sometimes referred to as insiders. Employees as insiders who are end users can be the weakest link causing risks to the organisation (Alohali, et al., 2018). With the increased demand to connect to cyberspace, most people and businesses use various end devices: smartphones, mobile devices, tablets, laptops, and computers (Imgraben, Engelbrecht, & Choo, 2014). Some organisations practice Bring Your Own Device (BYOD) as a policy that allows insiders to use their own devices such as external hard drives, laptops, and smartphones among others to access and exchange information (Ferdousi, 2022). The priviledge of employees using their own devices sometimes results in devices that get stolen, intruded upon, or information lost. Sometimes, employees perform certain activities on the Internet to attract criminals into the systems (McGuire & Dowling, 2013). Criminals are opportunists that are continually searching for the lucrative benefit and use any available and open interfaces on the network to gain unauthorised access (Ncubukezi, Mwansa, and Rocaries, 2020b).

End devices are one of the main resources that promote activities by both legitimate users and criminals in the business space (Ncubukezi, 2022b). The increased use of end devices introduces ways that compromise privacy, safety, and security in all organisations, requiring mitigation strategies and increased awareness (Mzileni & Ncubukezi, 2022). The literature shows that there is minimal understanding of users' safety practices on end devices in homes and business sectors (Tabassum, Kosinski & Lipford 2019). Employees ignorantly want to use their home devices at work, establish unsecured connections, pose a risk, and compromise computer and information security (Ncubukezi, 2022a).

Cyberattackers understand and pay close attention to the end devices as they become their primary targets for deploying malware attacks (Rusi & Lehto, 2017). Criminals steal information of good value and money by infiltrating end devices (Mosteanu, 2020). Likewise, these devices may not be connected to the business network but serve as standalone computers where employees share and exchange information. Regardless of the connection to the network, the end devices generally pose a risk because some end devices do not use proper security measures owing to the access of BYOD. In many different institutions, the culture of BYOD and lack of strong security on end devices can result in a range of cyber security risks such as highjacked unauthorized access

to the resources, loss, breach, and leakage of a data breach, regular phishing or malware attacks, poor handling of data (Bhadouria, A.S.; Ferdousi, 2022; Shukla et al., 2022).

Some end devices could experience power supply failure or a single point of failure where a device can crash or malfunction due to poor practice of device safety and regular protection of the devices (Ncubukezi, 2022b; Hong, 2017). Some could even display error messages that could freeze the device or automatically shut down owing to corrupt software or faulty hardware (Ncubukezi, Mwansa & Rocaries, 2020b). These risks related to the devices could cost the companies and requires the use of artificially intelligent systems to help organisations to make better security decisions and improve the security of the resources and data handling to reduce risks (Ncubukezi, Mwansa & Rocaries, 2020a).

After analyzing the relevant and available literature, this study determined the risk likelihood of the end devices using the Bayesian network tools. To achieve the main goal, the paper:

- Used scenario analysis and decision tree analysis to understand end device risks
- Illustrated and simulated end device risk likelihood and the impact using the Bayesian network
- Performed sensitivity analysis using Tornado graphs, conditional probability tables (CPT), and value of information configuration (VOI)

The rest of the paper is organized as follows: The method used to carry out the study is presented in the following section, followed by the sensitivity analysis results with discussions, and the conclusion of the study is accounted for.

1.1 Possible risks associated with end devices

End devices generally form part of the network infrastructure. These devices on the network can range from mobile devices, laptops, desktops, servers, phones, and printers and can be used for transmission for either the source or the destination. For data transmission, these devices get allocated a unique number that identifies them with a network address (Ncubukezi, Mwansa & Rocaries, 2020b). So, these devices can operate stand-alone or in a networked environment which exposes them to a wide range of risks. Table 1 shows the risk exposure of end devices, the form of attack, related security principles, the risk likelihood, and the risk impact. End devices are exposed to a variety of risks such as device loss, theft or robbery, damage due to third-party involvement, malware, denial of service or phishing as well as unauthorised access to both the facilities where resources are stored and the actual devices. These risks are initiated by employees or criminals in the form of actions performed by ignorant employees. At times, the insiders make poor decisions, computer illiterate on poorly configured devices due to a lack of policy enforcement and poor management (Ncubukezi, Mwansa & Rocaries, 2021). All these sources of device attacks challenge security principles such as confidentiality, integrity, and the availability of stored information on the devices (Mzileni & Ncubukezi, 2022).

According to Table 1, end devices are most likely to have software and hardware failure, misplacing old equipment, affected by constant network downtime, malware, phishing, or the denial of service as well as the negative activities performed by legitimate system users (Wheelus & Zhu, 2020). The impact of the end device risk probability affects the devices and information stored, organisations at large, and the current and potential clients (Ncubukezi, 2022b).

Table 1 which illustrates factors that contribute to the risk likelihood of the end devices is presented below.

Table 1: Risk of end devices (Ncubukezi, 2022b)

END DEVICES				
Risk exposure	**Form of attack**	**Related security principle**	**Risk likelihood**	**Risk Impact**
• Theft, loss or robbery • Third-party involvement • Network related (Malware attack or denial of service) • Unauthorised access to devices	• Human actions • Ignorance - poor decision-making • Lack of employee skills • No device encryption • Lack of device management	• Availability • Confidentiality • Integrity	• Hardware & software failure • Old equipment • Network downtime and denial of service Human factors • Network related events	• Poor production & business growth • Fraudulent acts • Lack of clients and their trust • Loss of intellectual property • Poor economic growth • Loss or stolen information and devices • Network related (denial of service, malware) • Delayed delivery of the services and economic growth • Hardware failure, slow or unusable device

1.2 Related works

There has been an increase in the use of information and communication technologies (ICT) for Internet and information processing which leaves institutions vulnerable to a diverse range of risks. The exposure of the different components of the business sectors to risks necessitates the determination of the risk likelihood using computers and other intelligent systems such as the Bayesian Network (BN). For example, a study conducted by Ncubukezi (2023) determined the risk probability of human-related actions within the business small sector to determine the risk probability. The study simulated the risk probability and impact based on the human factor variables (dependent and independent), and their relationships resulting in data leakage. On a cloud-based model, other authors used multilevel BN to assess the risk likelihood of gas transmission stations and present different flows (Gao et al., 2022). This study also used the BN to determine the risk likelihood on the end devices used for information exchange.

Even though institutions use IoT devices such as different gadgets, machines, sensors, and other hardware devices for information exchange, this study focuses on the end devices such as computers, printers, smartphones, and laptops. Devices can pose risks on different levels. For example, a more detailed scenario for IoT devices which is not part of the scope of this study is presented in the study conducted by Muñoz, Fernández-Gago, and López-Villa (2022). Their work described the risks and attacks on the protocol level. In addition, the use of hardware and software certifications of the end devices to improve security is also not part of the scope of the study but a study conducted by Munoz and Mafia (2014) has thoroughly presented and combined the model. These two studies complement this study in the sense that one work reveals the weaknesses of IoT devices at the protocol level and the other study proposes a certification solution using secure elements for establishing the pseudo-hardware certification system. This work uses different techniques with BN tools to determine the risk likelihood of the end devices.

The following section presents the method used to carry out the study.

2. Methodology

The Covid-19 global pandemic revealed an increase in cyber-attacks on all levels resulting in a range of risks within different business sectors. This study focuses on end device exploitation to determine the risk likelihood using the risk analysis methods and the Bayesian network. Risk analysis techniques such as scenario analysis and decision tree analysis (DTA) simulate the end device case and demonstrates the possibilities of the risk probability. They also provide an assessment of the likelihood that project decisions affect risk measures.

On the contrary, Bayesian Network tools are demonstrated using the AgenaRisk platform which illustrates the variables used and their relationships resulting in the risk likelihood and its impact. Sensitivity analysis was used to determine the risk probability of the end devices using the Tornado graphs, conditional probability tables (CPT), and value of information configuration (VOI). Sensitivity analysis is a quantitative technique that helps to determine variables that have a greater impact on risk (Cox, 2008). Variables estimate risks with potential impact and determine the changes in objectives and uncertainties that are correlated, along with the effect of each element on the objectives. It also communicates data and outcomes, understanding the link between the input and output variables while identifying sensitive variables (Wang & Neil, 2021; Khodakarami, & Abdi, 2014). It then helps to make assumptions that allow decision-making and examines the amount of risk in given scenarios. Only scenarios with the greatest risk are taken into account in the sensitivity analysis. The study also developed a different Bayesian model (BM) as part of the graphical probabilistic model that used the Bayesian inference analysis. The simulated model illustrates the risk likelihood and the risk impact of end devices. The use of these Bayesian network techniques determines the risk likelihood.

As cyber risks increased, they need a system with artificial intelligence capabilities that perform better than the human mind in predicting risks (Ncubukezi, Mwansa & Rocaries, 2020a). This work used the Bayesian Network as a model on the AgenaRisk package to predict the risk likelihood of the end devices. The BN helps to develop the probabilistic model by demonstrating and presenting different nodes with random variables to influence conditional probability for other dependent variables (Ramesh et al., 2003). BN in this study determines end device risk probability based on the prior indicators (networked or stand-alone), device protection levels, risk impact, and risk likelihood. The interface of the AgenaRisk package is shown in Figure 1 demonstrates different scenario cases of phishing attacks.

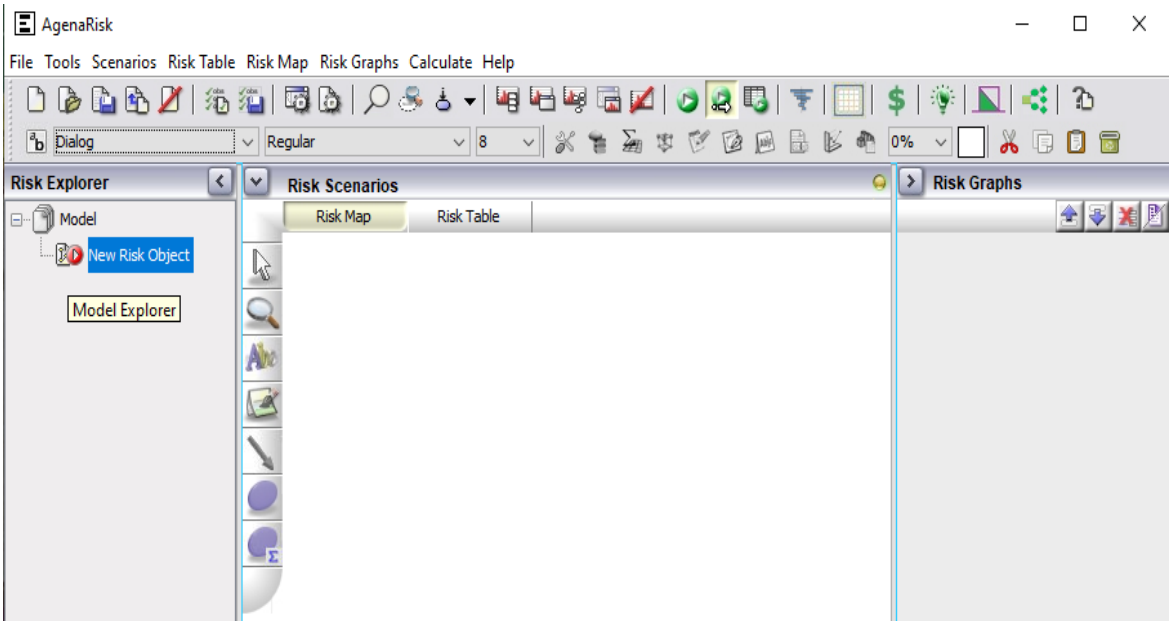


Figure 1: AgenaRisk interface

Ethical considerations: To conduct this work, the researcher received the ethical clearance that grants permission for the research to be conducted. The simulation of the end device scenario is presented below.

3. Simulation

This work presents the scenario of the end devices by using the decision tree analysis shown in Figure 2. The analysis describes how the end devices become vulnerable to risks. Two sources of risks experienced by the access level devices are mentioned.

3.1 Scenario and Decision Tree Analysis (DTA)

Devices are primarily used to penetrate and access business services. In the business sector, there are both networking devices (routers, switches, hubs), while the end devices can be smartphones, notebooks, laptops, and desktops. If these devices are connected to the network they become vulnerable to a range of cyber threats and attacks, which result in different risks to businesses (Ncubukezi, Mwansa and Rocaries, 2021). Figure 2 shows the cybercriminal and insider-related actions used to exploit the end devices. Within business institutions, the risk relating to the end devices can be caused by both the cyber criminals and insiders (employees) that take advantage. Cybercriminals as attackers gain unauthorised access to networked devices by using a piece of code to cause and trigger unintended behavior on the device. This process hijacks authorized users' actions by deploying a code that manipulates the device (Ncubukezi, 2022b).

With unsafe storage and handling of the end devices, cybercriminals also use opportunities to steal the devices and the information stored on the devices. On the other hand, criminals sometimes deploy software code to guess and steal passwords so that they may gain unauthorised access to business systems. These actions are performed by gaining unauthorised access through the network and result in poor and malfunctioning network performance in the form of denial of service or malware. Cybercriminals also hijack the network so that it may not be available and can further hijack user privileges (Ncubukezi, 2022a).

On the contrary, employees ignorantly perform activities that leave the business system vulnerable. Some employees could lose the device, which could land in the criminals' hands or they can be criminals by stealing the devices. Sometimes, the employees use unsecured memory sticks for information exchange. Those memory sticks could have malware that triggers device exploitation. In addition, using the Internet could result in device exploitation because employees could ignorantly download software from unknown sources. This scenario describes how the end devices can be exposed to risks. These activities are shown in Figure 2.

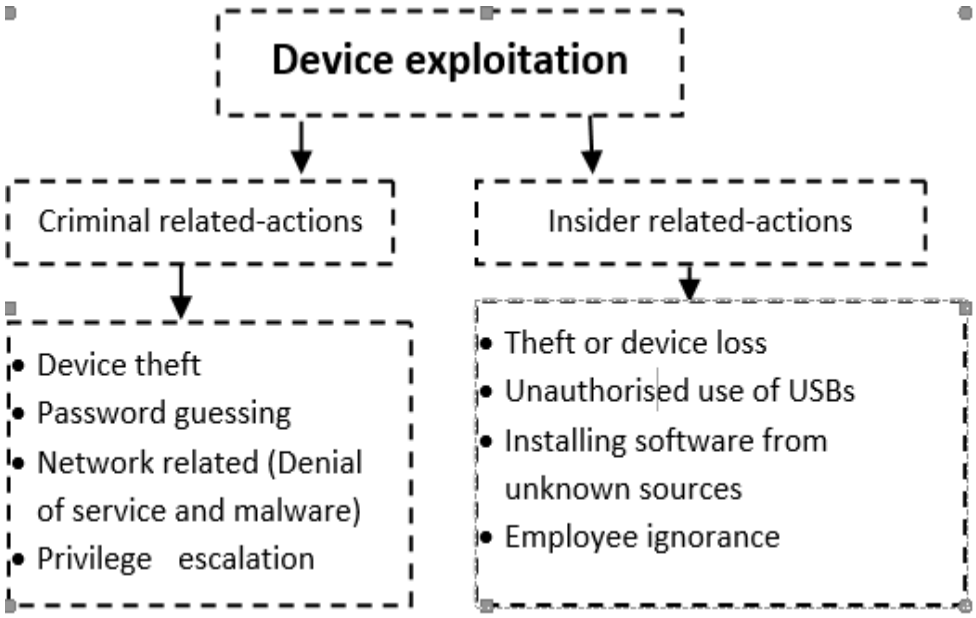


Figure 2: End Device exploitation

This study also used the Bayesian network tools with the AgenaRisk package to determine the potential cybersecurity risk probabilities and their impacts in the business space. BM is part of the graphical probabilistic models that use the Bayesian inference analysis (van de Schoot, et al., 2021). The BN analysis demonstrates the end devices' risk likelihood using sensitivity analysis techniques. The application of the BN is presented below.

3.2 Bayesian Network Model setting

The main variables of the end device scenarios are the prior indicators of the receiving end devices, the device protection level, and the posterior indicators, which influence the risk likelihood and impact. As used on the different models, the prior indicators are the end devices that could be either portable or fixed and connected to the network or act as standalone devices. The device's protection level is influenced by implementing security measures such as updated software (application and operating systems), device encryption, physical protection, and hardware security. In addition, complete device protection is also influenced by the nature of a device (networked or standalone). The level of device protection affects the posterior indicators to determine the risk likelihood as well as the impact. The higher the device protection, the lower the impact, even though the risk likelihood would be influenced by the device protection, portable or fixed device, which could be networked or standalone. This section demonstrates the sensitive analysis of the devices which result in different cyber risks. Sensitivity analysis challenges the reliability, difference, and significance of the assumptions to address the 'what if' analysis. In addition, sensitivity analysis is the quantitative technique for determining variables that have a greater impact on risk (Cox, 2008). Different end device scenarios are simulated and presented below.

3.3 End device risk likelihood simulation

The simulations of the end device scenario are based on different possible conditions. According to the figures, end-device simulations range from 1 to 7. The influencers of these scenarios are the state of an end device (fixed or portable), networked or stand-alone, and device protection level (software, device encryption, physical protection, and security of the hardware). These prior indicators influence and determines the risk likelihood and the risk impact of the device per scenario. These different scenarios are presented below.

Simulation 1: Figure 3 illustrates fixed and standalone devices with 84% high device-protection level based on 100% outdated software updates and 100% lack of hardware security. The device is in a secured physical environment with a 78% high level of encryption. Based on the overall device protection, this scenario had a medium-risk likelihood with 57% of the risk impact.

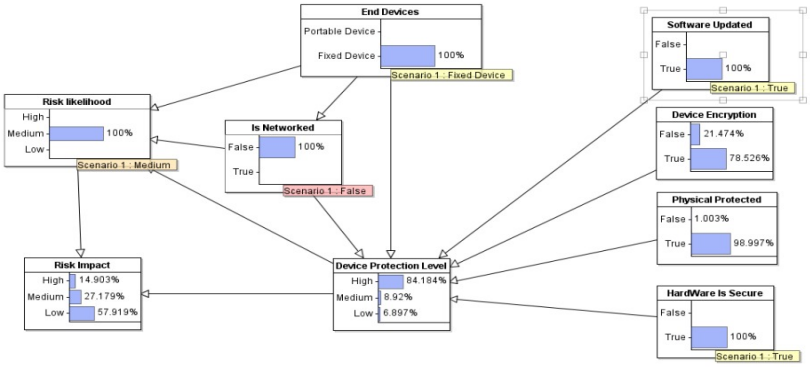


Figure 3: Fixed standalone end devices with high protection level

Simulation 2: Presents fixed standalone end devices with low hardware and software protection. Figure 4 illustrates networked fixed-end devices with a 41% low device-protection level. This scenario has no software updates, 80% high device encryption, 99% physical safety, and 100% hardware security. Inconsistent device protection influences medium risk likelihood and 36% medium risk impact.

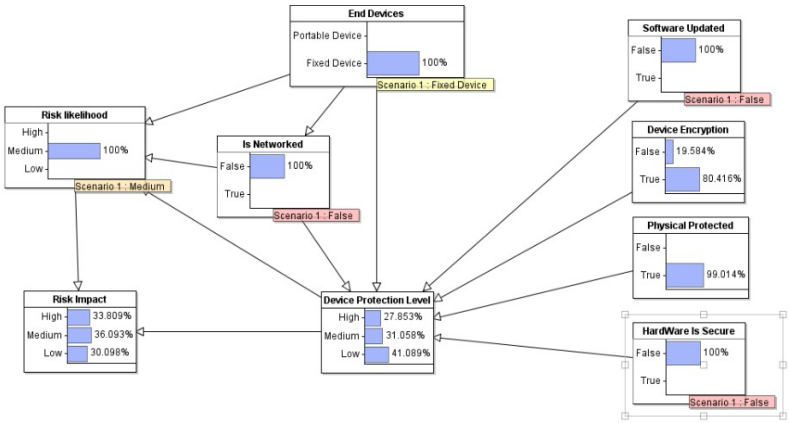


Figure 4: Fixed standalone end devices with low protection level

Simulation 3: Figure 5 presents portable networked end devices with high device protection levels owing to the physical space, hardware, and device encryption. The software becomes the only low influential factor for risk level. The range of the high mediating factors influences the device security's overall level, determining the risk likelihood and impact.

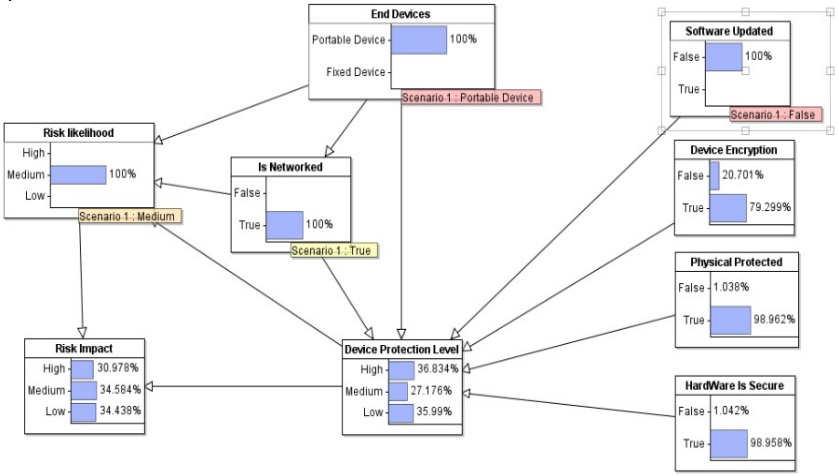


Figure 5: Portable networked end devices with high device protection

Simulation 4: On Figure 6 demonstrates a networked portable end device with a 57% low protection level owing to no hardware and software updates, 98% dominating physical protection, and 79% device encryption. The dominating 57% low device protection level has a medium-risk probability and 42% high-risk impact.

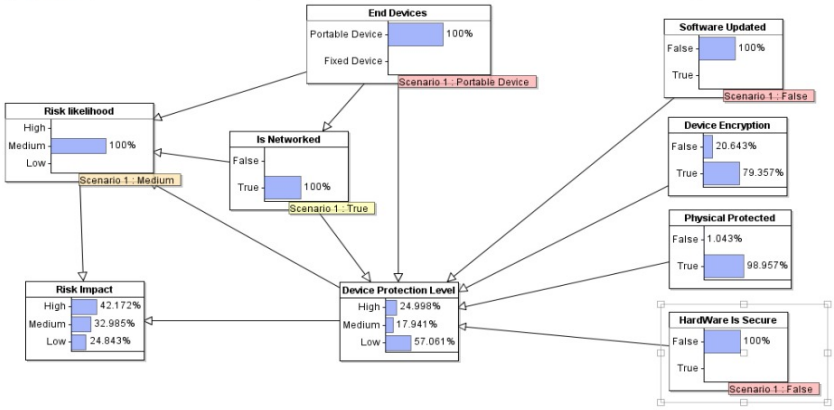


Figure 6: Portable networked end devices with low hardware and software security

Simulation 5: Figure 7 presents a networked portable end device scenario with medium-risk likelihood and 34% medium-risk impact. The 36% of device protection has influenced the results. In this scenario, the software is not regularly updated, only 79% of device encryption, 98% of physical safety of the buildings, and 98% of hardware security.

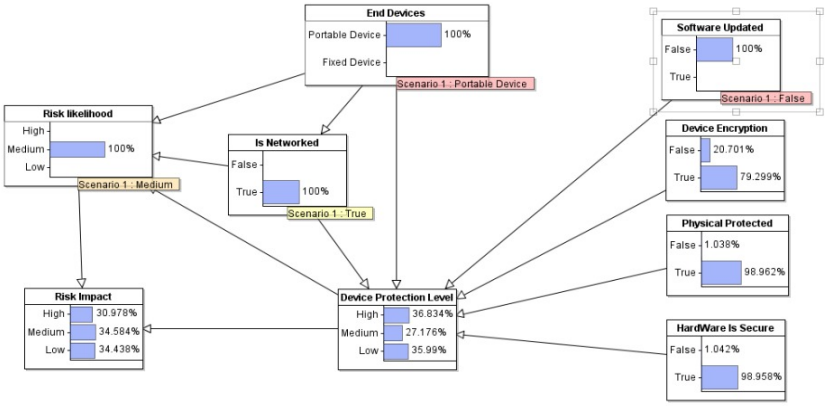


Figure 7: Portable networked end devices with low software protection

Simulation 6: Figure 8 shows the portable standalone end devices with an 84% protection level resulted in an 83% low-risk likelihood and a 75% low-risk impact. Safety measures are 99% software updates, 99% hardware, and 80% physical security, resulting in 84% device protection.

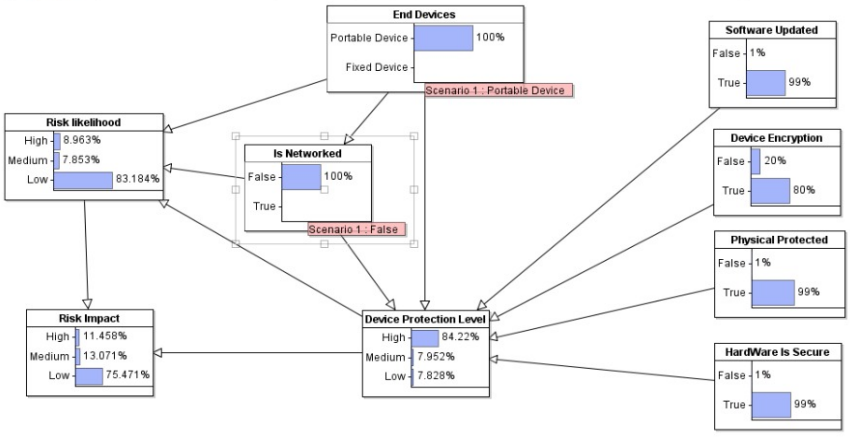


Figure 8: Portable standalone end devices with high device protection

Simulation 7: Figure 9 shows a 100% low device protection level due to 96% software updates, 22% device encryption, 98% physical protection, and 97% hardware protection. All the variables, in this case, contribute to the overall cyber risk likelihood and impact medium-risk likelihood results and 5% low-risk impact for a portable networked device.

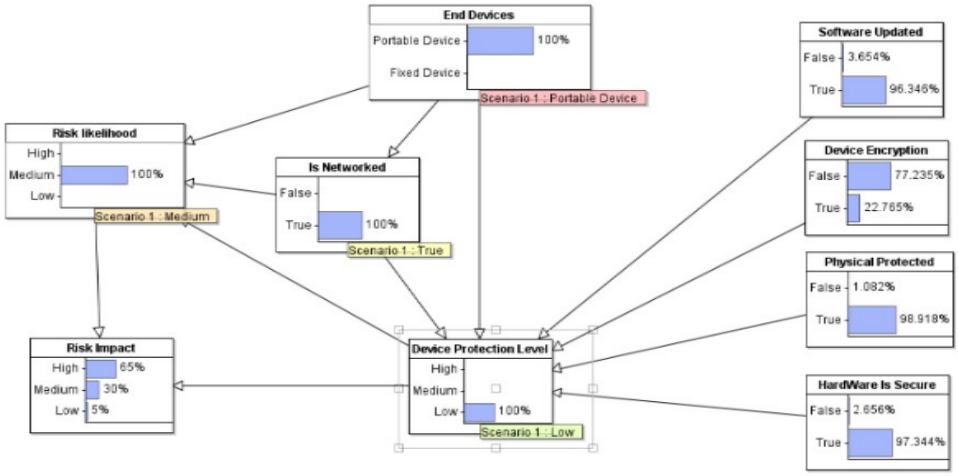


Figure 9: Portable networked end devices with low device protection level

4. Sensitivity Analysis Results

This part of the work presents the end device results based on the sensitivity analysis and the prediction of the risk probability using the Bayesian Network. This study used the AgenaRisk to conduct an extensive sensitivity analysis of the end device scenarios to check the sensitivity of the answers against the technique and its related parameters. The study is a hypothetical analysed sensitivity of the AgenaRisk package in isolation, in that if numerous interrelated parameters relate to the answer, then the researcher only considers the effect of one parameter at a time. For the tool's effectiveness, the researcher determined the technique's sensitivity to the Conditional Probability Tables (CPT). The sensitivity analysis communicates data and outcomes, understanding the link between the input and output variables while identifying sensitive variables. It then helps to make assumptions that allow decision-making and examines the amount of risk in given scenarios. The tornado graphs are generated based on the study's scenario analysis. Below are the CPT results followed by the Tornado graphs and the value of information configuration.

4.1 Conditional Probability Tables

The conditional probability tables (CPT) represent the probabilistic dependence of the parameters in BN models. The table shows results based on the observation outcomes which are generated based on the available prior knowledge. This section presents the conditional probability table for the end devices, which determined the sensitivity outcome based on the threat level that results in a data breach. CPT has the threat level ranked high, medium, and low, while the data breach has two Boolean states, which are true and false. Figure 10 shows the CPT for end devices with a false data breach when the threat level is high (0.007), medium (0.347), and low (0.646). The data breach is true when the threat level is high (0.631), medium (0.311), and low (0.058).

		Threat Level		
		High	Medium	Low
Data breach	False	0.007	0.347	0.646
	True	0.631	0.311	0.058

Figure 10: CPT for end devices

4.2 Tornado graphs

Tornado graphs present a chart that demonstrates the sensitivity based on given deciding variables. Each variable can have a different range of values which directly affects the impact and the risk likelihood. As a graphical representation of the sensitivity analysis, the tornado graph illustrates the sensitivity of the dependent and independent variables. On the AgenaRisk platform, the graph is generated by clicking on the Tornado chart button. The graph also presented different scenarios for sensitivity analysis. So this study presented three Tornado graphs for different rankings of the threat levels, which are high, medium, and low, and two Boolean values (true and false) for the data breach. The Tornado technique examined the sensitivity analysis for the end device-related risks. Figure 11 shows the Tornado graph for a false (0.007) data breach and (0.631) threat level.

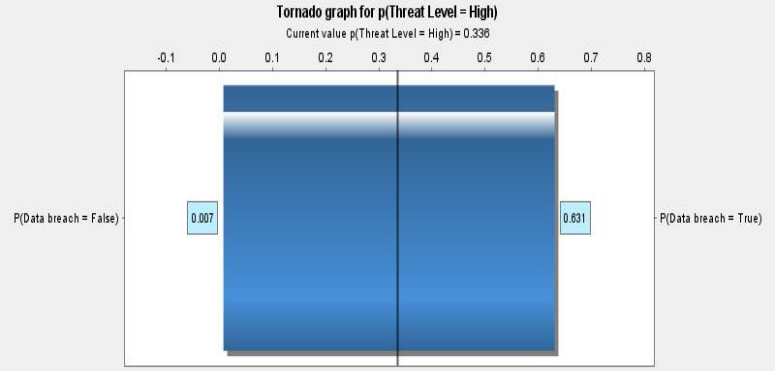


Figure 11: Tornado graph for false data breaches and high threats level

Figure 12 shows the Tornado graph for a true (0.311) data breach and (0.347) medium threat level.

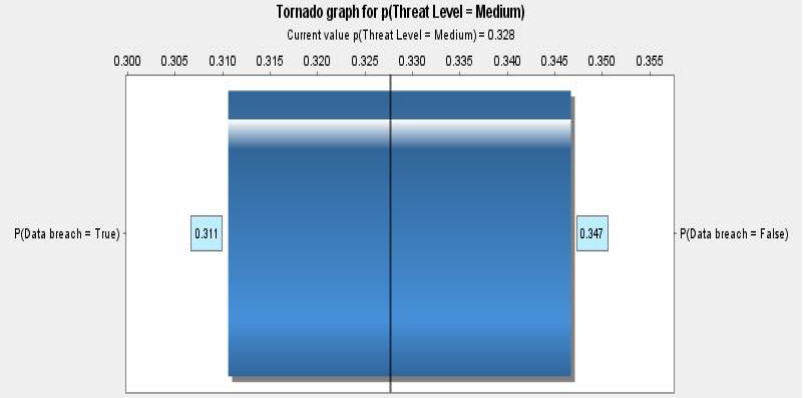


Figure 12: Tornado graph for true data breach and medium threats level

Figure 13 shows the Tornado graph for a true (0.058) data breach and (0648) low threat level.

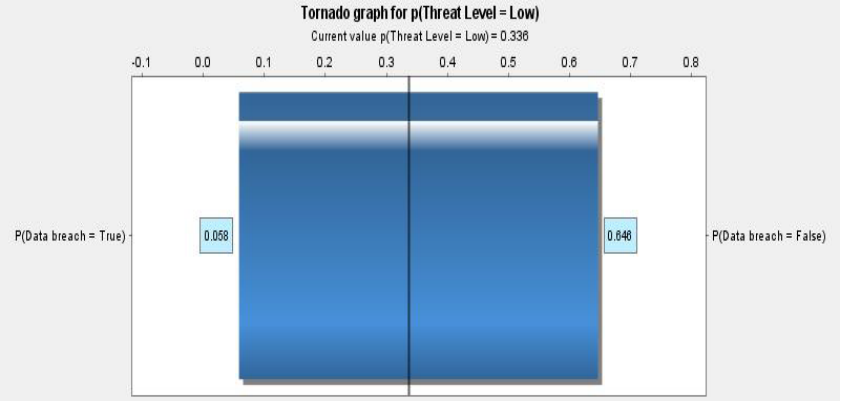


Figure 13: Tornado graph for true data breach and low threats level

4.3 Value of information configuration

The study used VOI as the sensitivity analytical method for decision-making based on the quantified available information to clarify difficult issues and reduce any risk uncertainties. In addition, VOI determines the sensitivity of the observations to improve decision-making. As used in the study, the value of information (VOI) configuration presents the amount that businesses or management could pay for required information that will help them to make decisions. This method is used to obtain more information that relates to the end device risk scenario that is used with the AgenaRisk package. The method reduces uncertainties prior decision-making stage. So, this work presented the decision nodes used in the scenario, the total time to build the risk scenarios in seconds, the expected maximum value (EMV), the expected value given perfect information (EV|PI), and the expected values of partially perfect information (EV(P)PI) for each configuration. Figure 14 shows the configuration for the devices and technical systems with the probability of risk as a decision node, the level of protection of the device as an uncertainty node, and the impact of risk as a utility node. This scenario took 122 minutes to perform the analysis of the expected maximum value (EMV), the expected value given perfect information (EV|PI), and the expected values of partially perfect information (EV(P)PI).

VOI Configuration	
Decision Node	Risk likelihood [M7]
Uncertainty Nodes	Device Protection Level [M1]
Utility Node	Risk Impact [M8]
Optimisation Type	maximum
Scenario	Scenario 1

Total build time: 122 ms

Expected Maximum Value (Utility|Decision) – EMV

Expected Value Given Perfect Information – EV|PI

Expected Value of (Partially) Perfect Information – EV(P)PI

Figure 14: Configuration information for end devices

As shown in Figure 15, the devices' EMV is 0.759 with two different malware conditions, EV|PI of 0.728 and EV(P)PI of -0.031. In addition, the device protection has three rankings of low, medium, and high, while the risk likelihood is ranked low, medium, and high.

EMV				**0.759**
Device Protection Level [M1]			**EV|PI**	0.728
				-0.031
Click to show/hide details				
		Risk likelihood		
		High	Medium	Low
Device Protection Level	High	0.65	0.68	**0.767**
	Medium	0.553	0.563	**0.683**
	Low	0.283	**0.3**	0.23
EV|PI = 0.859 * 0.767 + 0.071 * 0.683 + 0.07 * 0.3 = 0.728				
EV(P)PI = 0.728 - 0.759 = -0.031				

Figure 15: End devices' EMV

4.4 Discussions

This work determined the risk likelihood and the impact of the end devices with different capabilities. The end device simulation case scenario revealed different outcomes of the risk probability and impact, which are influenced by the prior indicators. Observations and the use of available data such as the prior indicators were used to determine sensitivity analysis using CPT, Tornado graphs, and VOI. CPT determined the end device risk likelihood based on the events of the different variables that result in a risk. In addition, CPT helped to determine the future events of the risks relating to the end devices. The input data was generated and computed to determine the risk outcome (Alkhairy et al., 2020). The results of using the Tornado graphs demonstrated the risk impact and the likelihood of the end devices based on the available prior indicators to determine the level

of the risk likelihood and extent of the risk impact. The results of the VOI showed the sensitivity analysis which could help to identify potential risks relating to the end devices. Organizations and other institutions can use the VOI outcomes to use advanced tools which can reduce costs relating to the end risks (Zabeo et al., 2019).

The analytical techniques used in the study revealed that the safety and security of the end devices depend on the different implementations of the security measures. These measures are linked to the physical hardware of the device, network connection, and software installed. The diverse contributing factors showed the varying probability of the risk related to the end devices. The minimal implementation of device protection security measures exposed the transmitting end devices to a range of cyber threats (Ncubukezi, Mwansa & Rocaries, 2020b).

Even though some end devices may be fixed and not connected to the network, the device's protection level continues to influence the risk level and its impact. The physical security of the devices may not reduce the risk probability of the device if the hardware, software, and encryption are ignored. The guarantee of the end devices should be applied at all levels (hardware, physical, software, and encryption) (Ncubukezi & Mwansa, 2021). If the device's security is prioritised for the hardware, software, physical, and encryption level, the level of risk to the device will be low, resulting in a low-risk impact. For good hygiene of the devices, a wide range of security controls, remedies, and measures should be used at different levels. For example, implementing firewalls, and intrusion detection systems (IDS) would be beneficial to securing and protecting the business's assets (Rawindaran et al., 2021).

Ncubukezi, Mwansa, and Rocaries (2020b) emphasize using adequate security measures to promote good cyber hygiene. With the high rate of risk uncertainties, it would be beneficial for the small business sector to deploy protective measures at all levels of end devices. Poor management of the institutional infrastructure including the end devices becomes the primary target of cyber attackers and results in most institutions losing their private and sensitive information. In addition, during the process of transmission, attackers also gain unauthorised access to the data that is transmitted. Furthermore, networking devices also become the victims that releasing confidential data (Lamba et al., 2017).

Recommendations: Several authors suggest the following safety practices that promote guide hygiene for end devices. Even though institutions differ, organisations need to take precautions and safety measures to improve the security of the devices, and the information stored on the devices. Ncubukezi (2022a), Ncubukezi and Mwansa (2021), Ncubukezi, Mwansa, and Rocaries (2020b) Williams, Chaturvedi, and Chakravarthy (2020), Sung et al., (2018) suggest the following recommendations to reduce the probability of risk on end devices.

- Perform regular system and software updates
- Use of the firewall to proactively filter traffic
- Use of antivirus, and antispyware software to protect the devices and the system
- Use of device encryption in case of the lost device
- Ensure devices are always on secure and locked storage
- Use multi-factor authentication to improve access to the device and its systems
- Perform regular software and hardware updates to reduce device downtime and failure
- Periodically back up data to avoid the single point of failure
- Ensure endpoint protection by securing the devices

5. Conclusion

This study determined the risk likelihood of the end devices using different analytical methods to predict the risk likelihood and their impact. As the end devices operate on the access level of every network, they become vulnerable to different risks. The sources of criminal and insider-related risks were accounted for. The work described the end device scenario which was interpreted using the decision tree analysis – which showed the possible risks devices are vulnerable to. In addition, BN tools were used to analyse sensitivity analysis of the end device risk. These tools include the VOI, CPT, and Tornado graphs. VOI showed the EMV of the device. CPT revealed the threat level on devices. Tornado graphs showed the threat levels for different threat rankings that results in the ultimate data breach.

The paper achieved the usage of the Bayesian network to illustrate the risk likelihood of the end device variables using the dependent and independent variables. The prior indicators used were the software updates,

encryption, physical security, and safety of the hardware determined the level of the risk probability and the risk impact. The outcomes produced different results which revealed low, medium, and high risks of the devices. The study benefits all institutions and people who use the devices to transmit data on the network to proactively take measures that protect and secure devices. In addition, management would come up with policies that will address and promote device security.

Acknowledgments

I would like to thank National Research Foundation (NRF) Black Academics Advancement Programme (BAAP) Grant for the time to conduct the study and financially supporting this work.

References

Alkhairy, I., Low-Choy, S., Murray, J., Wang, J. and Pettitt, A., 2020. Quantifying conditional probability tables in Bayesian networks: Bayesian regression for scenario-based encoding of elicited expert assessments on feral pig habitat. *Journal of Applied Statistics*, 47(10), pp.1848-1884.

Alohali, M., Clarke, N., Li, F. and Furnell, S., 2018. Identifying and predicting the factors affecting end-users' risk-taking behavior. *Information & Computer Security*.

Bhadouria, A.S., Study of: Impact of Malicious Attacks and Data Breach on the Growth and Performance of the Company and Few of the World's Biggest Data Breaches.

Cox, A. 2008. What's wrong with risk matrices? *Risk Analysis: An International Journal*, *28*(2), pp.497–512.

Ferdousi, B., 2022. Cyber security risks of bringing your own device (BYOD) practice in the workplace and strategies to address the risks. *International Journal of Science Academic Research.* 3(10), pp.4554-4558.

Gao, P., Li, W., Sun, Y. and Liu, S., 2022. Risk assessment for gas transmission station based on cloud model-based multilevel Bayesian network from the perspective of multi-flow intersecting theory. *Process Safety and Environmental Protection*, *159*, pp.887-898.

Hong, H.J., 2017. From cloud computing to fog computing: unleash the power of edge and end devices. In 2017 IEEE *International Conference on Cloud Computing Technology and Science (CloudCom)* (pp. 331-334). IEEE.

Imgraben, J., Engelbrecht, A. and Choo, K.K.R. 2014. Always connected, but are smart mobile users getting more security savvy? A survey of smart mobile device users. *Behaviour & Information Technology*, 33(12), pp. 1347–1360.

Khodakarami, V. and Abdi, A., 2014. Project cost risk analysis: A Bayesian networks approach for modeling dependencies between cost items. *International Journal of Project Management*, *32*(7), pp.1233-1245.

Lamba, A., Singh, S., Balvinder, S., Dutta, N. and Rela, S., 2017. Analyzing and fixing cyber security threats for supply chain management. *In International Journal For Technological Research In Engineer*ing, 4(5).

McGuire, M. and Dowling, S., 2013. Cybercrime: A review of the evidence. Summary of key findings and implications. *Home Office Research report*, 75, pp.1-35.

Mosteanu, N.R., 2020. Artificial Intelligence and Cyber Security–A Shield against Cyberattack as a Risk Business Management Tool–Case of European Countries. *Quality-Access to Success*, *21*(175).

Munoz, A. and Mafia, A., 2014. Software and hardware certification techniques in a combined certification model. In *2014 11th International Conference on Security and Cryptography (SECRYPT)* (pp. 1-6). IEEE.

Muñoz, A., Fernández-Gago, C. and López-Villa, R., 2022. A Test Environment for Wireless Hacking in Domestic IoT Scenarios. *Mobile Networks and Applications*, pp.1-10.

Mzileni, I., Ncubukezi, T., 2022. Impact of information security threats on small businesses during the Covid-19 pandemic. In *European Conference on Cyber Warfare and Securit*y, 21(1), pp. 401-410.

Ncubukezi, T. and Mwansa, L., 2021. Best Practices Used by Businesses to Maintain Good Cyber Hygiene During Covid-19 Pandemic. *Journal of Internet Technology and Secured Transactions*, 9(1), pp.714-721.

Ncubukezi, T., 2022a. Human Errors: A Cybersecurity Concern and the Weakest Link to Small Businesses. Conference Proceedings: *International Conference on Cyber Warfare and Security,* 17(1), pp. 395–403.

Ncubukezi, T., 2022b. *Design development and evaluation of the cybersecurity risk tool: a case of small and medium-sized enterprises in South Africa.* (Doctoral Thesis, Cape Peninsula University of Technology, unpublished).

Ncubukezi, T., 2023. Risk Likelihood of Planned and Unplanned Cyber-Attacks in Small Business Sectors: A Cybersecurity Concern. In *International Conference on Cyber Warfare and Security* (pp. 279-XV). Academic Conferences International Limited.

Ncubukezi, T., Mwansa, L. and Rocaries, F., 2020a. A Proposed: Integration of the Monte Carlo model and the Bayes network to Propose Cyber Security Risk Assessment Tool for Small and Medium Enterprises in South Africa. *International Journal of Computer Science and Information Security*, 3(18), pp.152-155.

Ncubukezi, T., Mwansa, L. and Rocaries, F., 2020b. A review of the current cyber hygiene in small and medium-sized businesses. Conference Proceedings: *International Conference for Internet Technology and Secured Transactions (ICITST*), 15, pp. 1–6, IEEE.

Ncubukezi, T., Mwansa, L. and Rocaries, F., 2021. Analysis and impact of the cybercrimes in the Western Cape small and medium-sized businesses. *Conference Proceedings: International Conference on Cyber Warfare and Security*, 16, pp. 425–235.

Ramesh, R., Mannan, M.A., Poo, A.N. and Keerthi, S.S., 2003. Thermal error measurement and modelling in machine tools. Part II. Hybrid Bayesian Network—support vector machine model. *International Journal of Machine Tools and Manufacture*, *43*(4), pp.405-419.

Rawindaran, N., Jayal, A., Prakash, E. and Hewage, C., 2021. Cost Benefits of Using Machine Learning Features in NIDS for Cyber Security in UK Small Medium Enterprises (SME). *Future Internet*, *13*(8), p.186.

Rusi and Lehto, 2017, T. and Lehto, M. 2017. Cyber threats megatrends in cyberspace. Conference proceedings: *International Conference on Management Leadership and Governance.* Academic Conferences and Publishing Limited, 5, pp.323.

Shukla, S., George, J.P., Tiwari, K. and Kureethara, J.V., 2022. Data Security. In *Data Ethics and Challenges* (pp. 41-59). Singapore: Springer Singapore.

Sung, W.J., Ahn, H.G., Kim, J.B. and Choi, S.G., 2018. Protecting end-device from replay attack on LoRaWAN. In 2018 20th *International Conference on Advanced Communication Technology (ICACT*) (pp.167-171).

Tabassum, M., Kosinski, T. and Lipford, H.R. 2019. I don't own the data: End-user perceptions of smart home device data practices and risks. *Conference proceedings: Symposium on Usable Privacy and Security (SOUPS)*, 15, pp.435–450.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M.G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J. and Yau, C., 2021. Bayesian statistics and modeling. *Nature Reviews Methods Primers*, *1*(1), pp.1-26.

Wang, J. and Neil, M., 2021. A Bayesian-network-based cybersecurity adversarial risk analysis framework with numerical examples. *arXiv preprint arXiv:2106.00471*.

Wheelus, C. and Zhu, X., 2020. IoT network security: Threats, risks, and a data-driven defense framework. IoT, 1(2), pp.259-285.

Williams, C.M., Chaturvedi, R. and Chakravarthy, K., 2020. Cybersecurity risks in a pandemic. *Journal of medical Internet research*, *22*(9), p.e23692.

Zabeo, A., Keisler, J.M., Hristozov, D., Marcomini, A. and Linkov, I., 2019. Value of information analysis for assessing risks and benefits of nanotechnology innovation. *Environmental Sciences Europe*, *31*(1), pp.1-8.

Enabling fine-grained Access Control in Information Sharing with Structured data Formats

Tatu Niskanen and Jarno Salonen
VTT Technical Research Centre of Finland, Espoo, Finland

tatu.niskanen@vtt.fi
jarno.salonen@vtt.fi

Abstract: The ongoing need for societal and industrial digital transformation requires rapidly expanding networks of interconnected organizations and dictates an increasing role for cybersecurity in information sharing. A typical setup consists of multiple stakeholders working closely together and needing efficient channels for sharing relevant information in a secure manner. This is especially prevalent with complex modern supply chains and critical information infrastructures. They often comprise of numerous co-operating organizations, people and in some cases smart devices having different levels of access to a variety of information. Granular access control plays a vital role when distributing information efficiently between stakeholders without revealing sensitive pieces of data to unwanted third parties. This article presents a novel framework for enabling fine-grained access control to share information efficiently and securely in these situations. Our motivation and use case for the framework originates from the secure sharing of cyber incident information in the maritime logistics industry. We present a novel solution to this problem by developing an information sharing platform and a meta-model, demonstrated using an implementation with structured JSON data formats, while supporting previously researched attribute-based encryption schemes. The proposed framework provides a broader context to the fine-grained data access control challenge in addition to the technical implementation.

Keywords: Cybersecurity, Information security, Access control, Information sharing, Incident management, Resilience.

1. Introduction

Widespread and complex supply chains are common in the modern world of industry 4.0. This is evident for example in the automotive industry, where a single product has countless parts with separate manufacturers embedded into it. The supply chains and stakeholder networks surrounding these products are complex and require well refined measures to share all the necessary information securely with everyone involved. This is also evident with critical information infrastructures (CIIs). CIIs usually have complex architectures composed of many interconnected components with personnel having access to different levels of information. This article presents a novel framework to enable fine-grained access control (FGAC) for secure and efficient information sharing in situations like these ones.

FGAC systems facilitate granting differential access rights to a set of users and allow flexibility in specifying the access rights of individual users (Goyal et al, 2006). FGAC allows these access rights to be implemented using higher granularity than what traditional folder or file-level access control makes possible. This becomes relevant when we need to share a bundle of information, like a manual or a log file securely and efficiently with many stakeholders, while considering the different authorization of each individual to the specific data points. With a well-refined FGAC solution, we can share the entire manual or log file as such to the stakeholders and control read and write access to the different sections for everyone separately.

We produced a practical use case for our framework from sharing cyber incident information among stakeholders in the maritime logistics industry and related supply chains. An increasing number of complex cyber-attacks has affected the maritime sector, which is somewhat typical for modern industries relying heavily on cyber-connected systems. In order to combat this, Jacq et al. (2021) have identified a need for increasing the cyber risk awareness of the industry's smart ports.

In order to achieve this vision of increased cyber risk awareness, it is vital to define appropriate mechanisms for sharing and managing of threat and incident information. With our proposed solution, organizations can share incident data quickly and efficiently, while allowing access to the individual data points only to authorized actors. We validate our framework in a case-study implementation to a maritime cybersecurity exercise environment. Although our model allows for FGAC in sharing of cyber incident information, it is not limited to this type of information. We can extend our meta-model to most other types of information one wishes to share in their operational environment.

Our research contribution is a practical framework that organizations can implement to enable FGAC in information sharing. It utilizes a novel and flexible solution that can be applied to any data format which uses name/value pairs. We validate the framework with a case-study. During our research, we also discovered that

the extant literature has to some degree been very centred around the technical implementation and has thus partially left out the surrounding context regarding implementing FGAC in real-life environments. Therefore, in addition to the novel technical solution, our research also addresses some important issues regarding this surrounding context. Our research questions are the following:

1. How can we enable FGAC in sharing of sensitive information?
2. How can we create structured data formats which enable granular encryption and decryption?
3. How can we create a practically applicable framework which organizations can use to enable FGAC in information sharing?

We have structured the article as follows. After a brief introduction to the topic in this chapter, we describe the relevant theoretical background in chapter two. In chapter three, we present our framework and the four processes that enable its' implementation. In chapter four, we implement the framework in a case-study targeted to the maritime environment. Finally, we discuss about the implementation and conclude the article in chapter five.

2. Theoretical background

In this section we present the most noteworthy findings from current literature and research regarding our application of FGAC. We discuss about alternative FGAC implementations, relevant access control and encryption schemes, and some previous research concerning information sharing with structured data formats.

Access control is the process of mediating every request to resources and data maintained by a system and determining whether the request should be granted or denied (Samarati et al, 2000). Different access control policies can be applied, which correspond to different criteria on what is allowed. These policies form a basis for many different access control models, which represent how an organization enforces access control in practice, for example in a computer system. The different access control models are a well-researched domain and understanding them forms the basis for applying FGAC in information sharing. Some of the most implemented and researched access control models include role-based access control (Sandhu, 1998), attribute-based access control (Federal CIO Council, 2009 and Hu et al, 2015), and organization-based access control (Kalam et al., 2003). We further elaborate these models in the next section, alongside discussing the appropriate underlying access control models for applying FGAC.

The enforcement of access control often depends on the underlying encryption schemes. Some schemes are more suitable than others, or even specifically designed to carry out FGAC. An example of an encryption scheme designed to achieve FGAC is the work done by Goyal et al (2006). This encryption scheme called attribute-based encryption (ABE) works by labelling ciphertexts with a set of attributes. Cryptographic decryption keys are then associated with access structures that control which ciphertexts the user can decrypt. This approach however has some drawbacks, mainly related to the lack of a straightforward attribute and key revocation mechanisms.

FGAC has been the subject of research in different contexts. Some research has focused on the application of FGAC in the cloud environment. The work by Wang et al. (2010) presents a solution to achieve FGAC with the previously mentioned ABE schemes in the cloud environment. FGAC has also been implemented using blockchain solutions. Wang et al. (2018) present a blockchain-based framework to achieve FGAC and mention that the private key generator has the ability to decrypt all data stored in the cloud server in all ABE schemes, which may result in serious problems such as key abuse and privacy data leakage. They aim to provide a blockchain-based solution, which solves the problem of a single point of failure with a decentralized storage model. Blockchain however also has some disadvantages, and it may not always be the preferred solution by the implementer. The disadvantages of blockchain implementations include for example high transaction costs and difficult integration especially with older systems. (Golosova et al, 2018.)

With regards to the act of sharing information, especially considering our motivation of applying FGAC to the sharing of cyber incident information, some key background can be found in the RFC-standards. RFC 5070 defines the incident object description exchange format (IODEF). This is a standardized format which is used to share computer security information between computer security incident response teams. (Danyliw et al, 2007.) IODEF version 2 specified in RFC 7970 introduces new information points to this format (Danyliw, 2016). Salonen et al. (2022) have introduced a somewhat primitive model to enhance the sharing of cyber incident information via FGAC, utilizing IODEF to achieve this goal. This article provides a more mature version of that model and extends it beyond IODEF and incident information sharing.

3. Designing the Framework

We designed a framework to achieve FGAC for granular information sharing, with the purpose of presenting a replicable solution to this problem. We constructed the framework according to the following process:

1. Establishing and verifying user roles/attributes.
2. Setting up an information sharing channel.
3. Establishing a structured format for the information type.
4. Distributing access rights to the shared information.

In this section, we explain these parts and describe them in further detail. The first step makes the future work less demanding by providing a clear picture of the users' access control requirements in the implementer's operational environment. The second part is about setting up the venue through which we handle information sharing and apply FGAC. In the third part, a structured format is established for a given information type. This predefined format is needed to make the FGAC solutions of this framework implementable. The fourth section handles the software solutions, which allow for distribution of access rights, and the encryption methods.

3.1 Establishing and verifying user roles/attributes

In order to succeed in applying FGAC securely and efficiently, we must first plan and verify appropriate user roles or attributes in the access control framework of the target environment. This is vital for formulating a clear picture of who requires access to which information and becomes critical later when granular access rights are distributed to the shared information. The proper implementation of this step causes less manual work in later steps.

Multi-organization environments are the most relevant targets for FGAC. Establishing a solid access control policy and managing all the individual access rights can be challenging when multiple organizations operate concurrently with differing systems, personnel, and access control models. Some access control models are specifically designed to make handling this task more manageable, such as the Multi-OrBAC (Abou El Kalam et al, 2006) and PolyOrBAC (Abou El Kalam et al, 2009) models. They build on the foundations of organization-based access control (OrBAC) (Kalam et al, 2003). These applications are worth exploring, and the chosen access control model allows a freedom of choice for the implementer. However, for the purposes of applying this framework, a role-based or attribute-based solution is sufficient. The priority is that the chosen model gives us a clear picture of the different roles or attributes that the users have, which we can base our access control on.

Role-based access control (RBAC) is a policy-neutral model based on roles and role permissions (Sandhu, 1998). RBAC is known to have some pitfalls, which should be acknowledged. RBAC is not always well suited for provisioning all access control happening in a complex organization, since this can quickly lead to an explosion in the number of roles. For this reason, it should only be applied in applications which it is well suited for. In the more complex cases, RBAC may prove incapable of specifying users' access rights, since specifying access control based on the user's role alone may prove inadequate. It is also important to understand that RBAC is an ongoing effort. As organizations change, role models must be reviewed and updated to match. Even though RBAC is known to have its limitations, it is still a valid solution in certain applications. (Ferraiolo et al, 2001)

Attribute-based access control (ABAC) is a model where a subject's authorization to perform a set of operations is based on attributes associated with the subject, object, requested operations, and in some cases environment attributes (Hu et al, 2015). It is sometimes seen as the successor to RBAC. This framework is well suited for an ABAC implementation, although we demonstrate it using only the role-attribute in our case-study. An ABAC implementation does allow for further specifying of access rights using other attributes. We elaborate on this in the fourth step where we discuss the technical implementation.

3.2 Setting up an information sharing channel

In order to handle the sharing of documents, along with enabling their granular encryption and decryption, we must establish an information sharing channel. This makes implementing FGAC manageable in the scope of this framework. The channel acts as an intermediary between the information distributor and the consumer. It can be set up using a software-as-a-service implementation. Using this service, we can implement the software/encryption solutions required to apply FGAC to the shared information. We also use this service to view and manage the information, although this could be conducted locally by the users. The service needs to allow for the storage of:

- Encrypted documents
- User data to manage authorization
- Cryptographic keys for appropriate decryption of information

It is largely up to the implementer how they integrate user data into the service, as long as the users signing in have appropriate roles or attributes for managing access control. Figure 1 illustrates a possible service architecture.

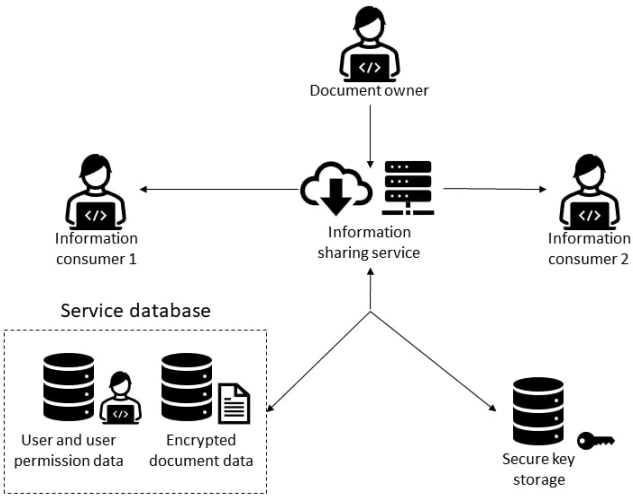


Figure 1 Illustration of a possible information sharing service architecture

The data should be stored in a secure database in an encrypted format. Our proposed solution is a MongoDB-database due to the flexibility its' NoSQL structure offers for managing JSON data. Depending on the chosen encryption scheme, the decryption will either be performed with securely stored document keys by the service, or the users' private keys using an ABE solution. We return to the different options for encryption in the fourth step.

The cryptographic keys should always be stored in a separate location from the encrypted data, as recommended by the Open Web Application Security Project (OWASP, 2022). This can mean storing them separately in the same system, but it is worthwhile to consider storing the keys on a separate system altogether. It is a known fact that the secure storage of cryptographic keys is a difficult problem to solve when the application needs to have some level of access to the keys in order to decrypt the data. OWASP recommends additional secure storage mechanisms for further security. These include a physical or a virtual hardware security module (HSM), key vaults, and secure storage APIs. (OWASP, 2022)

3.3 Establishing a structured format for the information type

In this part, a structured data format is created for the information to be shared. We use it to implement the encryption solutions which enable FGAC. The organization or multi-organization environment may want to come to agreement on a standardized format for the information type when possible. This eases the creation of subsequent documents of the same type.

The recommended data format in this framework is the JSON data format. JSON supports multiple individual fields and sub-fields using name/value pairs, which suits the methods of applying FGAC in this framework. The syntax and hierarchical structure of JSON strings make them easily interpretable by applications, and JSON can be parsed into a ready-to-use JavaScript object. As simple text, JSON is also suitable and safe for transferring across platforms and operating systems that do not share more complex document types. It is a format designed specifically for data interchange, outperforming alternatives such as XML in terms of speed and memory consumption. (Zunke et al, 2014)

For demonstration purposes, we assume that the shared information is the operations manual for a system component in a manufacturing environment. We can then construct a structured JSON format for that manual, resembling the IODEF data format. Table 1 clarifies this with an example. We apply this same method to our incident information sharing use case.

Table 1 Structured data format for a component manual

Field	Multiplicity	Authorization	Description
ComponentID	One	A4, B2, C6	A component identification number assigned to this component by the document creator
PartList	One	A4, C6	List of the parts contained in the component
ComponentDescription	One	B2	Description of the component
AssemblyManual	One	B2, C6	Manual to assemble the component
Contact	One or more	A4, B2, C6	Contact information to inquire further information
AdditionalData	Zero or one	A4, B2	Additional relevant data about the component

The table presents an example of a predetermined and structured data format. The authorization tags can correspond to user roles such as "maintenance worker", or other user attributes, which are used for access control management. Writing these attributes as corresponding tags instead of clear text descriptions adds an extra layer of security alongside a simpler presentation. The fields and possible sub-fields should clearly hold one type of information with a clear and concise authorization classification. When utilizing JSON, this format can be supported with a JSON-schema.

3.4 Distributing access rights to the shared information

This workflow step handles the embedding of access rights into the shared documents, along with the encryption methods and software solutions used in order to successfully manage FGAC.

After we have defined the authorization levels of the individual fields, the encryption process is performed. Enabling FGAC can be achieved by distributing one or more identifiers called access control tags to the individual fields. These correspond to user roles or other user attributes. We base these tags on user roles in our implementation. Access control tags remain as plaintext when we encrypt the document and bring it to the database, which allows for easier verification of privileges. The JSON data format does not support inserting images directly, but there are workarounds for this. One possible solution is to encode the images as ASCII text using Base64 encoding. Figure 2 displays the embedding of access control tags within a JSON file.

```
"ComponentId": {
  "roleTags": [
  "a4",
  "b3",
  "c4"
  ],
  "content": "f474a"
},
"PartList": {
  "roleTags": [
  "a4",
  "c2"
  ],
  "content": "Part 1, part 2, part 3."
},
```

Figure 2 Distributed access control tags in a JSON file

The distributed role-tags in this example correspond to the user roles which require access to the encrypted content within the field. All users registered to the service hold one or more roles. For example in a manufacturing environment, the distributed role-tag "a4" can correspond to a maintenance worker in a certain factory, whose role is indicated by a corresponding attribute. The permission to access/decrypt this field is then granted to the users of this role. With an ABAC solution it is also possible to implement tags here that correspond to other attributes besides roles. Figure 3 demonstrates an ABAC implementation:

```
"ComponentId": {
  "projectTags": [
  "a1",
  "b2",
  "c5"
  ],
  "locationTags": [
  "f7",
  "g3"
  ],
  "content": "f474a"
},
```

Figure 3 ABAC implementation of access control tags

We identify two suitable methods to handle the encryption using this solution:

- Symmetric-key server-side encryption and decryption.
- Attribute-based encryption schemes.

With the symmetric-key option, the implemented software solution is responsible for appropriate decryption of information for the users. We can check the access control tags under the JSON-fields directly against user roles or other attributes during decryption requests. Figure 4 demonstrates this in JavaScript.

```
if requestingUserTags.some(r =>
documentContent[field].tags.includes(r))
```

Figure 4 Verifying privileges in a JavaScript implementation

The encrypted information within the field can be decrypted for the user by the server if this requirement is fulfilled. The server fetches a corresponding document key for decryption from a secure key storage.

We base the second method of handling the encryption on previous work done by Goyal et al (2006). Applying ABE with our model requires that the underlying software solution labels ciphertexts with the appropriate attributes during the encryption/upload process, based on the access control tags implemented under the JSON fields. The software then generates corresponding private keys and distributes them to the users, enabling an asymmetric key scheme. The demonstration of this technical implementation will be future research.

4. Implementing the framework: a case-study

We developed a proof-of-concept information sharing service that demonstrates the framework in a practical setting. The use-case is based on the secure sharing of host-based and network monitoring data in maritime environment cybersecurity exercises. During Capture-the-Flag (CTF) type incident response exercises, we share network monitoring data captured during a simulated incident to the participants. The participants are granted access to the monitoring data according to their role and privileges in the context of the exercise. Thus, some of the participants are initially given limited access to the monitoring data, with only certain data points accessible. The data is used to formulate a situational picture of what has happened in the incident. The participants can be granted access to additional data points within the monitoring data, if they are initially unable to gain an adequate understanding of the situation.

The implementation of the service utilizes the methodologies presented in this framework. *We initiate the information sharing process of our case-study according to the first workflow step by defining what roles the participants have within the exercise*. We define two distinct roles; some participants are given managerial roles with more privileges, while others are left with basic roles with more limited access to the data. *As the second step of our framework dictates, we set up an information sharing channel to enable FGAC in the sharing of data to the participants*. Figure 5 illustrates the implemented information sharing service architecture in this case study:

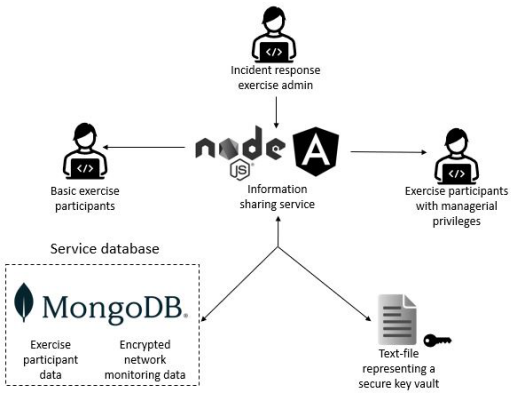


Figure 5 Implemented service architecture

The platform front-end was developed with the Angular 2+ web application framework for TypeScript, and the back-end solution was built with JavaScript's Node.JS framework. The tag-based encryption and the document-specific key generation is handled by the back-end solution, while the participants use the front-end interface to view the incident data during the exercise. We save the encrypted monitoring data and exercise participant data to a MongoDB-database, while we use a basic text file to save the keys. Although a text file in this situation is able to represent a secure key storage, we prefer other solutions like proper secure key vaults with a HSM in a real-world setting. The encryption is performed with the Advanced Encryption Standard using JavaScript's CryptoJS-library.

As instructed by the third step of our framework, we establish a structured format for the shared information. The network monitoring data for the exercise is polled with the Elasticsearch search engine. Elasticsearch is a suitable tool for our model since it can produce a JSON file of the entire network monitoring data, which can be uploaded to the service. Therefore, in this case we can produce our structured data format using the Elasticsearch search engine in a ready-to-use JSON-format. As the Elasticsearch query determines the fields and their content, we only need to note it down and define the authorization levels of each field for the users. With the roles defined in the first step, we use the indicator "u1" for basic users and "u2" for users with heightened privileges.

As the final step of our framework, *we distribute access rights to the data*. As an intermediary operation, we input the appropriate access control tags under the fields of the JSON file before the exercise admin uploads it to the service. We conduct this process manually in this case, but we could implement automated software solutions for this step in a relatively straightforward fashion. The encryption process is initiated when the document owner, in this case the exercise administrator, uploads the JSON file to the service. Regarding the chosen approach to encryption, we chose the symmetric-key option instead of ABE schemes. The software encrypts the field contents, while the tags and title remain as cleartext. The encrypted file is then saved to the MongoDB-database, as partially shown in Figure 6. A corresponding document key is saved to a secure key storage.


```
_id: 1
▾ fields: Array
  ▾ 0: Object
      content: "U2FsdGVkX18U0J5CU+AdXknA971Yhw1a56V2qIe3wsk="
    ▾ tags: Array
        0: "u1"
        1: "u2"
      title: "Incident-ID"
  ▾ 1: Object
      content: "U2FsdGVkX1+7E3OccuBGuzlytl7aSRCc5YstTqkqcnCGH7tDZ2y0ngUWrVcvUTbIYcFLwS…"
    ▾ tags: Array
        0: "u1"
        1: "u2"
      title: "Description"
  ▾ 2: Object
      content: "U2FsdGVkX1/BGGUNp1EIj27+A68UUYL/fstjbvkp7n8KclIsprYJismer3nXSwRj"
    ▾ tags: Array
        0: "u2"
      title: "Agent-ID"
```


Figure 6 JSON file parsed and encrypted into the database

With all the steps performed, we are able to display the implemented front-end for the case study users as our end result in Figure 7. It shows how a user is logged in as a basic participant with limited access to the data points, dictated by the implemented FGAC solution. When exercise participants access the service, they see a list of all the files that include tags which correspond to them. The decryption process is initiated when participants request to view a file by clicking it. The software goes through the document fields and decrypts data based on the access control tags that match the participant attributes within their user data.

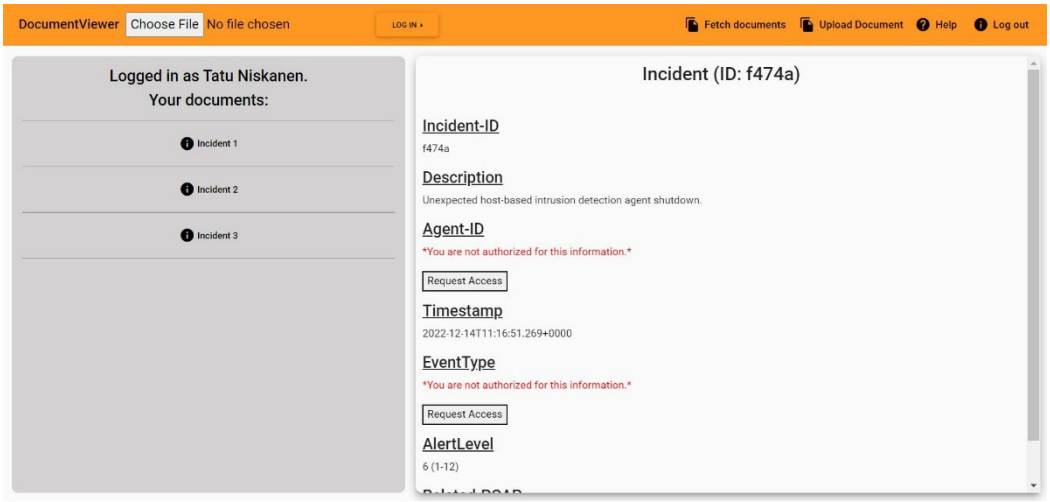


Figure 7 View of the front-end implementation with limited privileges

5. Discussion and conclusion

This article presents a practical solution to applying fine-grained access control (FGAC) in information sharing. The solution is based on a novel framework, motivated by information security needs identified within the maritime logistics and manufacturing industries. We have enabled it by using structured data formats and adding data specific access control tags to the individual data contents, while providing a broader context around the technical implementation.

By adopting the framework, organizations can enable FGAC in sensitive information sharing. The framework consists of four steps ranging from the establishment of user roles/attributes and setting up a suitable information-sharing channel, to establishing a structured data format and distributing access rights to the shared information. We describe a case study to demonstrate our framework and the practical use of FGAC in information sharing. In the case of utilizing this framework within another use case, the first workflow step needs to be implemented to fit the specific needs of the operational environment. The second step may be used as presented, although it can be altered to suit additional technical requirements. The third and fourth steps are in general re-usable in the presented form.

We provide an example on the establishment of a structured data format that enables granular encryption and decryption. Our example is based on the JSON data format, but the framework is compatible with any data format that supports name/value pairs. The framework supports multiple different encryption methods. Although we demonstrated the utility of this framework with the user roles as the determining attributes, other attributes could be used in their place. This can mean for example projects and locations that the users are assigned to.

Our future research and development activities consist of a more comprehensive piloting and evaluation of the service that can be done, e.g., in some other suitable future project. This would consist of among others a trial with actual users and real incident data that is added into the service and used for analysis. During the pilot, we could evaluate the service efficiency and applicability, as well as conduct surveys to collect end-users' views concerning the advantages and disadvantages of the service and system compared to other existing services and tools focused on incident management. Our future work also includes thorough security testing of the presented method. Another interesting topic for piloting would be evaluating the use and feasibility of alternative encryption mechanisms and/or data formats.

According to our research, the extant literature on FGAC has to some degree been mainly focused on the technical implementation and not so much on the surrounding context such as viable business cases. Therefore, we propose that more research should be conducted on the non-technical perspective and especially in realistic use cases such as the Cyber-MAR project (Jacq et al, 2021). Our proposed framework provides a template on how organizations can enable FGAC in information sharing. The future objective is to further validate the framework using real scenarios and to compare it more thoroughly with other potential competitors, adopting best practices and learning from their weaknesses.

Acknowledgements

This article is based on research and development work conducted in two Horizon 2020 projects, namely Secure Collaborative Intelligent Industrial Assets (SeCoIIA) and Cyber preparedness actions for a holistic approach and awareness raising in the MARitime logistics supply chain (Cyber-MAR). SeCoIIA aims at securing the digital transition of manufacturing industry towards more connected, collaborative, flexible and automated production techniques. Cyber-MAR aims to develop an innovative cybersecurity simulation environment for accommodating the peculiarities of the maritime sector, while being easily applicable in other transport subsectors, with the view to fully unlock the value of the use of cyber range in the maritime logistics value chain. The projects have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871967 and No. 833389.

References

Abou El Kalam, A. and Deswarte, Y. (2006). Multi-orbac: A new access control model for distributed, heterogeneous and collaborative systems. In Proceedings of the IEEE Symposium on Systems and Information Security.

Abou El Kalam, A., Deswarte, Y., Baïna, A., and Kaâniche, M. (2009). Polyorbac: A security framework for crit- ical infrastructures. International Journal of Critical Infrastructure Protection, 2(4):154–169.

Danyliw, R. (2016). The incident object description exchange format version 2. RFC 7970, RFC Editor.

Ferraiolo, D. F., Sandhu, R., Gavrila, S., Kuhn, D. R., and Chandramouli, R. (2001). Proposed nist standard for role-based access control. ACM Transactions on In- formation and System Security (TISSEC), 4(3):224– 274.

Golosova, J. and Romanovs, A. (2018). The advantages and disadvantages of the blockchain technology. In 2018 IEEE 6th workshop on advances in information, electronic and electrical engineering (AIEEE), pages 1–6. IEEE.

Goyal, V., Pandey, O., Sahai, A., and Waters, B. (2006). Attribute-based encryption for fine-grained access control of encrypted data. In Proceedings of the 13th ACM conference on Computer and communications security, pages 89–98.

Federal CIO Council (2009) Federal Identity, Credential, and Access Management (FICAM) Roadmap and 1671 Implementation Guidance Version 1.0, November 10, 2009.

Hu, V. C., Kuhn, D. R., Ferraiolo, D. F., and Voas, J. (2015). Attribute-based access control. Computer, 48(2):85– 88.

Jacq, O., Salazar, P. G., Parasuraman, K., Kuusijärvi, J., Gkaniatsou, A., Latsa, E., and Amditis, A. (2021). The cyber-mar project: First results and perspectives on the use of hybrid cyber ranges for port cyber risk assessment. In 2021 IEEE International Conference on Cyber Security and Resilience (CSR), pages 409– 414. IEEE.

Kalam, A. A. E., Baida, R. E., Balbiani, P., Benferhat, S., Cuppens, F., Deswarte, Y., Miege, A., Saurel, C., and Trouessin, G. (2003). Organization based access con- trol. In Proceedings POLICY 2003. IEEE 4th Interna- tional Workshop on Policies for Distributed Systems and Networks, pages 120–131. IEEE.

OWASP (2022). Owasp cheatsheets. https: //cheatsheetseries.owasp.org/cheatsheets/Cryptographic Storage Cheat Sheet.html. Accessed: 2022-08-15.

R. Danyliw, J. Meijer, Y. D. (2007). The incident object description exchange format. RFC 5070, RFC Editor.

Salonen, J., Niskanen, T., and Raitio, P. (2022). How to enhance the sharing of cyber incident information via fine-grained access control. In 3rd International Conference on Data Mining and Machine Learning, DMML 2022, page 17. AIRCC Publishing Corporation.

Samarati, P. and Vimercati, S. C. d. (2000). Access control: Policies, models, and mechanisms. In International School on Foundations of Security Analysis and De- sign, pages 137–196. Springer.

Sandhu, R. S. (1998). Role-based access control. In Advances in computers, volume 46, pages 237–286. Elsevier.

Wang, G., Liu, Q., and Wu, J. (2010). Hierarchical attribute- based encryption for fine-grained access control in cloud storage services. In Proceedings of the 17th ACM conference on Computer and communications security, pages 735–737.

Wang, S., Zhang, Y., and Zhang, Y. (2018). A blockchain- based framework for data sharing with fine-grained access control in decentralized storage systems. Ieee Access, 6:38437–38450.

Zunke, S. and D'Souza, V. (2014). Json vs xml: A comparative performance analysis of data exchange formats. IJCSN International Journal of Computer Science and Network, 3(4):257–261.

Towards an Active Cyber Defence Framework for SMMEs in Developing Countries

Nombeko Ntingi, Jaco du Toit and Sebastian von Solms
Academy of Computer Science and Software Engineering, University of Johannesburg, South Africa

nntingi@gmail.com
jacodt@uj.ac.za
basievs@uj.ac.za

Abstract: Small, medium, and micro enterprises (SMMEs) are obliged to adopt digital technologies to render services to their clients and remain competitive. The COVID-19 global crisis has accelerated the cyberfication of systems and services. The move to digital platforms has afforded SMMEs opportunities to offer their services to a broader geographical area. However, this has also presented opportunities for cybercriminals to invade the digital infrastructure. Adopting digital transformation has put SMMEs in a vulnerable position since they need to manage their cybersecurity while lacking the necessary skills and ICT infrastructure. The inability of SMMEs to defend themselves against cyberattacks compels them to outsource their security needs to external security service providers. These external security service providers offer security services based on a hierarchical operating model. Essential security services are offered at a lower level. If the paying clients require advanced security services, they may be provided as an add-on to the contractual agreement resulting in additional cost. This paper explores the active cyber defence (ACD) approach to enhance cybersecurity defence while minimising service costs. Therefore, the primary objective and outcome of this paper are to identify some of the essential drivers that will contribute towards developing the active cyber defence framework for SMMEs in developing countries. For purposes of clarity, essential drivers are the gaps highlighted during the literature review and will be referred to as "essential drivers" throughout the paper. The essential drivers, together with suggested recommendations, will be consolidated. The essential drivers were drawn from existing literature by going through peer-reviewed academic papers and company whitepapers. To achieve the primary objective, we need to establish whether SMMEs are utilising the services of external security service providers. The external security service providers will be referred to as "Security Operation Centre - SOC as a service" throughout the paper. The secondary objective of this paper is to determine whether SMMEs are utilising the SOC as a service and if they do, whether they realise value for money.

Keywords: SMMEs, Active Cyber Defence, Proactive, Artificial Intelligence, SOC as a Service, Security Service Provider

1. Introduction

According to a report conducted by Abbosh & Bissell (2019), organisations are more dependent on the digital economy and the internet than they have ever been in the last ten years. Large and small enterprises are obliged to adopt the digital economy to render their businesses efficiently and effectively to realise business value while expanding their client base. These enterprises are affected by cyber risks, such as reputational, financial, and regulatory consequences, irrespective of their sizes, geographical locations, or industries (Accenture/Poneman Institute, 2019). This is due to the ever-changing landscape of the cybersecurity domain. SMMEs are the softer target and easier to penetrate due to the inefficiency of their current cybersecurity defences (Weiss & Muegge, 2019). Cyber defenders face a huge task as they need to close all loopholes in ICT systems, whereas cybercriminals only need one weak point to penetrate through. It is not a myth that cybercriminals are always one step ahead of cyber defences and that cyber defences must always play catch-up. As cyberattacks become more sophisticated, more innovative, and proactive methods of hardening cyber defences are required.

When dealing with advanced or determined cybercriminals, such as state or state-sponsored attackers, the current passive cyber defence approaches are no longer enough, as they are just one part of the solution (Broeders, 2021). States and non-state organisations must bolster their defence mechanisms through proactive measures. One concept that nation-state and non-state organisations can use to complement passive cyber defences is Active Cyber Defence (ACD) (Burke, 2020).

This paper's primary objective is to identify some of the essential drivers that will contribute towards developing the active cyber defence framework for SMMEs within the same industry, in developing countries. To achieve the primary objective, we will also determine whether the SMMEs are utilising the SOC as a service and whether they are realising value for money.

The rest of the paper is organised as follows: Section 2 conceptualises active cyber defence. Section 3 gives background literature on the cybersecurity landscape of SMMEs in developing countries. Section 4 discusses the role of SMMEs in the economies of developing countries. Section 5 discusses the advanced cybersecurity solution –Security Operation Centres (SOCs), as implemented in large enterprises. Section 5 also discusses the

SOC as a service and highlights the SOC operating models. Section 6 discusses the optimisation of ACD using Artificial Intelligence (AI). Section 7 consolidates essential drivers that have been identified and suggest recommendations. Section 8 concludes the paper and highlights observations for future studies.

2. Conceptualisation of 'Active cyber defence'

In this section, we will briefly discuss the concept of active cyber defence (ACD) and its originality. According to Denning & Strawser (2017), ACD originates from the military or nation-state organisations, and it is derived from the active air and missile defence from the US Department of Defence. "Active air and missile defence is defined as the direct defensive action taken to destroy, nullify, or reduce the effectiveness of air and missile threats against friendly forces and assets" (Denning & Strawser, 2017).

The views on ACD are conflicting, with some scholars perceiving ACD as an intrusive concept that engages in hacking back. A comparative study on the US, Israeli and Germany states that ACD is an actively deployed countermeasure, with the countermeasures conducted during the ACD operations categorised as *defence with a twist, hack-back and persistent engagement and defending forward* (Herpig, Morgus & Sheniak, 2020). The *defence with a twist* countermeasure involves utilising defence security tools, with some tools having active and evasive capabilities (Herpig, *et al*., 2020). The *hack back* countermeasure involves using invasive tools to trace back the origins of the adversary with the intention of retaliating (Herpig, *et al*., 2020). The *persistent engagement and defending forward* involves using cyber capabilities to act even before the adversary attacks, disrupting the threat's capabilities from its source (Herpig*, et al*., 2020). From this comparative study, it can be deduced that ACD involves both offensive and defensive operations. However, Broeders (2021) defines active cyber defence as a concept that allows organisations to go beyond passive defences by implementing proactive cyber defence mechanisms for their own ICT networks. Organisations are not allowed to trace back where the cyberattack originated from or retaliate against cyber attackers beyond the perimeters of their own ICT networks (Broeders,2021).

As illustrated in Figure 1 below, by dotted lines on the active-defensive quadrant, the ACD falls on the active-defensive quadrant of the cyber counterintelligence (CCI) model (Ntingi, Duvenage, du Toit & von Solms, 2022). Even though the tools and techniques used to conduct activities of active cyber offence and active cyber defence are the same, active cyber defence aims to strengthen the cyber defences of an organisation rather than intrusive actions towards the adversaries (Duvenage, Jaquire & von Solms., 2020a; Ntingi, *et al*., 2022).

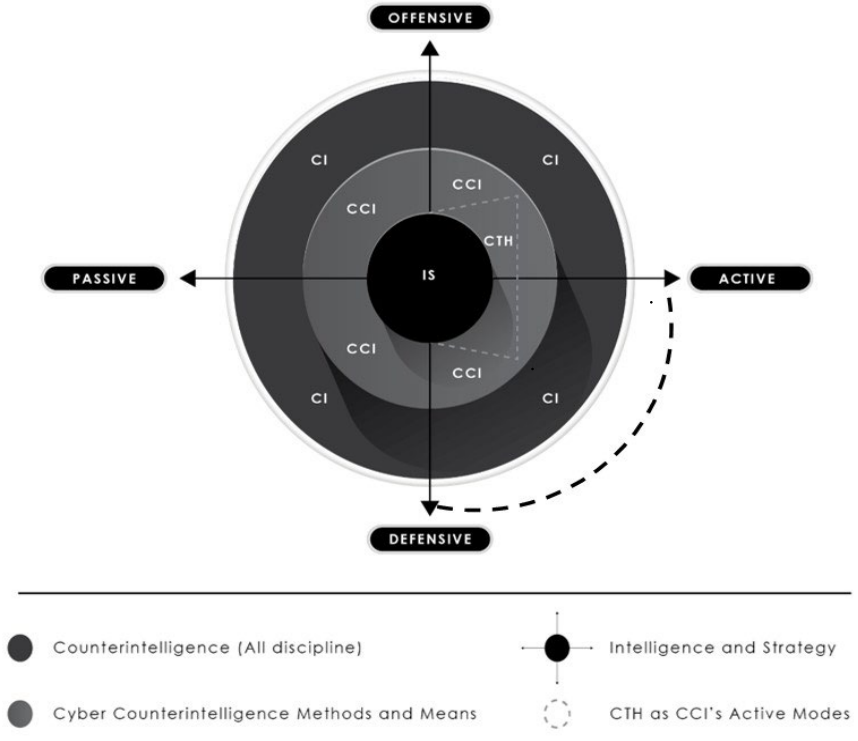


Figure 1: Cyber Threat Hunting as part of Cyber Counterintelligence (Ntingi, *et al*., 2022)

Our focus is on the cybersecurity operations of organisations, particularly SMMEs, to strengthen their cyber defences and actively defend their ICT systems without resorting to intrusive actions against adversaries.

Having discussed the active cyber defence concept, we will now discuss the cybersecurity landscape of developing countries.

3. Cybersecurity landscape of SMMEs in the developing countries

To be able to achieve the primary objective of the paper, which is to identify the essential drivers that will contribute towards the development of the active cyber defence framework, it is imperative that we foreground that with a brief discussion of the cybersecurity landscape of SMMEs in developing countries. This background information will highlight and strengthen the need to develop an ACD framework for SMMEs within the same industry, in developing countries. Organisations of all sizes, large and small, are taking advantage of the benefits that come with internet usage because it broadness the scope of service delivery. Digital transformation allows organisations to compete on a global scale and reduce the turnaround time of service delivery. According to a report by Verizon (2019), 43% of SMMEs in South Africa were victims of security breaches. A review conducted by Ncubukezi, Mwansa & Rocaries (2020) on SMMEs in the Western Cape, South Africa, states that SMMEs do not prioritise cybersecurity risk assessment and analysis. The review further states that there are no unanimous cybersecurity best practices across various industries (Ncubukezi, *et al*., 2020). Consequently, this leads to different cybersecurity approaches used to maintain cybersecurity hygiene (Ncubukezi, *et al*., 2020).

A study conducted on South African SMMEs by Kabanda Tanner & Kent (2018) reveals that the organisational internal factors, such as budget, management support, and attitudes, hinder the perception towards cybersecurity implementation. Consequently, these factors are regarded as impediments to implementing cybersecurity in SMMEs'environment (Kabanda *et al*., 2018). Additionally, SMMEs lack the know-how to identify cyber breaches, which causes delays in responding to cyber breaches, worsening the impact and adversely affecting business continuity.

While the cybersecurity state of SMMEs in developing countries is dire, a study by Mimecast (2022) reveals that organisations are becoming more aware of the need to implement and strengthen their cyber defences. There is a growing recognition that a lack of cyber preparedness is a major risk factor (Mimecast, 2022). 96% of the respondents reported that their organisations either already have or are developing cyber defence strategies (Mimecast, 2022).

The following subsection will group the essential drivers that were identified in this section.

3.1 Essential drivers identified in section 3.

These essential drivers are the identified challenges confronting SMMEs in developing countries. The essential drivers have been identified in section 3 above.

- Different cybersecurity approaches are used to maintain cybersecurity hygiene.
- Lack of financial resources.
- Limited or lack of advanced cybersecurity skillsets.

Flowing from the discussion, it is evident that even though SMMEs are willing to embrace cyber defences, there are still challenges confronting them, such as a limited budget and a lack of necessary skill sets to carry out cyber defence operations. The following section will discuss the role of SMMEs in the economies of developing countries.

4. The role of SMMEs in the economy of the developing countries

The role of SMMEs in the economies of developing countries cannot be understated. SMMEs are considered economic drivers and vehicles for job creation. Statistics South Africa, StatsSA (2019) estimated a labour force of about 14,7 million in the third quarter of 2020, with SMMEs accounting for approximately 45% of employment. Furthermore, SMMEs contributed 39% of the turnover of South Africa's gross domestic product (GDP) in the first quarter of 2019 (StatsSA, 2019). Additionally, SMMEs are part of the supply chain as they act as third-party contractors and provide services to large organisations. Considering the role of SMMEs in the economies of the countries, to stay competitive and relevant in the borderless economies and advancement of technology, they are compelled to embrace technology to offer their services efficiently to their clients. Adopting technology by SMMEs is a critical enabler in expanding business and client engagement. However, the adoption

of digital transformation brings with it cybersecurity challenges. While digital transformation has many benefits and opportunities in the ways in which organisations conduct their businesses, it also exposes them to significant cybersecurity risks.

The ever-changing landscape of the cybersecurity domain is faced with evolving cyber risks. The principal cyber risks that face organisations, as stated by the Institute of Risk Management (2014.), include business operational risk, reputational risk, and legal and compliance risk. SMMEs being part of the supply chain may present opportunities as the weakest link for cybercriminals to violate the supply chain and attack large organisations (de Vicente Mohino, Mallouli, Ruiz & van Haastrecht, 2021). Due to the exposure of SMMEs to cyberattacks, it is prevalent for SMMEs to adopt emerging proactive approaches towards the safeguarding of ICT infrastructure.

It is against this background that there is a need for SMMEs to implement and strengthen cyber defence technologies tailored to meet their specific needs. The following section will briefly discuss the technology solution, known as Security Operation Centres (SOCs), as currently implemented in large enterprises. The SMMEs can also adopt the SOC, but there is a need to adapt it to suit SMMEs' environments.

5. Security operation centres (SOCs) - as implemented in large enterprises

This section will briefly discuss the SOC as implemented in large enterprises. Furthermore, the SOC as a service and its operating models will be discussed. The concept of community SOC will also be introduced in subsection 5.3. Briefly, a community SOC concept services a group of companies in the same industry. The objective of this section is not to highlight all the SOC operating models but to highlight some of the models and find gaps that can be enhanced for the specific needs of SMMEs.

The SOC is defined as a centralised organisational unit whose primary functions are to continuously detect, analyse and respond to cybersecurity threats (Vielberth, Bohm, Fichtinger & Pernul, 2020; National Institute of Standards and Technology, 2018). Many organisations are building SOCs to combat the actions of adversaries. The function of the SOC is accomplished using a combination of people, technology, and processes (Vielberth, *et al*., 2020). SOCs in large enterprises can also implement advanced functions, such as threat hunting, incident analysis and response to highly sophisticated threats. These advanced functions are regarded as optional in the SOC environment due to their complexity and the high levels of expertise required (Kaspersky, 2019). SOCs are large and complex, necessitating large budgets, advanced skillsets, and advanced technologies. The SOC can either be operated internally within the organisation or externally through the SOC as a service. According to a survey conducted by Crowley & Pescatore (2018), enterprises that have built their own SOCs are confronted by various challenges in establishing well-run efficient SOCs. Some of the challenges include a lack of skilled staff, a lack of automation, a lack of integrated tools, and being overwhelmed by too many alerts (Crowley & Pescatore,2018). Due to SMMEs being just as vulnerable to cyberattacks as large enterprises and yet lack the necessary resources and skillsets to defend themselves against these cyberattacks, they are compelled to outsource security functions to SOC as a service (Weiss & Muegge,2019; Gupta & Zhdanov, 2012). Mihindu & Khosrow-shahi (2020) also asserts that with constraints confronting SMMEs, such as limited budget, SMMEs need a new and perhaps a different approach towards implementing cyber defences.

Having discussed the fundamentals of the SOC, we will now discuss the SOC as a service. The discussion will aid in establishing whether SMMEs fully utilise these cybersecurity services and, if not, highlight the factors that act as barriers to that.

5.1 SOC as a service

This section will briefly discuss the SOC as a service and highlight its operating models that can be adapted to better suit the cybersecurity needs of SMMEs in developing countries. According to Harrison (2020), to run a 24 x 7 SOC, organisations need to employ at least a minimum of 5 full-time professionals. Due to budgetary constraints and a lack of skills, enterprises, including large and small, have resorted to outsourcing their cybersecurity needs to third-party organisations. SOC as a service is, therefore, an answer to resource-constrained enterprises.

SOC as a service follows an operating model when delivering its services to its paying clients. In this section, we will discuss managed security service providers (MSSP) and managed detection and response providers (MDR) operating models. The role of these external service providers differs in how they offer their services to their paying clients. Both MSSPs and MDRs deliver security functions to their clients based on the contractual

agreement. According to TechTarget (2021), MSSPs often provide basic SOC services such as security management and monitoring services. This means that if a cybersecurity breach or intrusion is detected, the cybersecurity provider only alerts the client. Unless specified on the contractual agreement as an add-on, the advanced cybersecurity functions will be at an additional cost.

According to Zimmerman (2014); Vielberth, *et al*., (2020), the operating model is categorised into three tiers. The most basic services are offered at the first tier, with advanced cybersecurity services offered at the higher-level tiers. Red Canary Webinars (2022) also states that MSSPs offer their services using a tiered approach, with tier 1 services as basic services and tier 2 and 3 services being advanced security services offered at an additional cost. In contrast, MDRs are specialised services that offer more advanced and complex security functions to their paying clients (TechTarget, 2021). It is worth noting that the delivery of SOC services is scattered, hence a need for consolidation of the current operating models.

Now that we have discussed the operating model of the SOC as a service, in subsection 5.2, we will continue to address the secondary objective. In the interest of clarity, we will re-emphasise our secondary objective.

For us to be able to achieve this primary objective, we need to investigate the secondary objective.

- The secondary objective of this paper is to investigate whether SMMEs are utilising the SOC as a service and whether they realise value for money.

5.2 Utilisation of SOC as a service by the SMMEs

The aim of this section is to investigate the utilisation of SOC as a service by SMMEs and the challenges confronting them in optimally utilising SOC services. One of the challenges facing enterprises in implementing SOC as a service is that the service providers may use different security tools that cannot be integrated with each other (Kaspersky, 2019). Furthermore, these service providers might not support certain applications used in SMMEs. If the application is not part of the list supported by the external service provider, then SMMEs might have to consider procuring additional ICT infrastructure.

As mentioned in subsection 5.1, the delivery of SOC services is scattered. For enterprises to fully enjoy the benefits of SOC as a service, they may need to combine the services of MSSPs and MDRs to supplement the security functions of the other. Some functionalities offered by these services may be irrelevant to the needs of SMMEs and may therefore escalate the cost of cybersecurity implementation. For SOC as a service to deliver its services, some organisations might require significant changes in sourcing additional ICT infrastructure (Schatz, 2014). Organisations would be forced into "locked-in", resulting in rising costs of switching to other service providers, thereby compelling them to use the cybersecurity service providers even when their security needs are no longer met (Schatz, 2014).

The challenges mentioned in this subsection are some of the major impediments to SMMEs not utilising the

SOC as a service. It is for this reason that developing a community SOC framework for SMMEs in developing countries is imperative. Having highlighted the impediments to utilising SOC as a service, we will now introduce the concept of a community SOC.

5.3 Conceptualising a community SOC

In this section, we will introduce the concept of a community SOC. According to a survey conducted by Abbosh & Bissell (2019) on building trust in the digital economy, 86% of respondents thought that in the near future, a code of conduct for businesses in the same industry would be necessary to build resilience and foster trust among them. Based on this notion, developing a community SOC framework for SMMEs in the same industry would be feasible and assist in reducing the cost of implementing active cyber defences. The idea of a community SOC is to provide services to SMMEs from a specific community in the same industry and serve as a central and intermediate SOC that feeds into a higher-level SOC.

The UK government promotes the creation of community SOCs, termed Warning, Advice and Reporting Point (WARP), for public sector communities. The WARPs are made up of members that have a collective interest, such as the same business sector or geographical area. The responsibility of the WARPs is to act as a single point of contact and provide information security warnings and advice services to communities in the same industry (NLAWARP, 2018). For example, the defence industry WARP acts as a single point of contact for cyber incidents that pose risks, specifically to the Ministry of Defence. The critical cyber incidents are escalated from the local WARP to a higher-level cyber centre with advanced capacity (NLAWARP, 2018).

In South Africa, the National Cybersecurity Policy Framework (NCPF) recommends the establishment of relevant additional computer security response teams (CSIRTs) that will serve as a point of contact for that specific sector for cybersecurity matters, as well as coordinate cybersecurity incident response actions for that sector (State Security Agency, 2015). NCPF encourages collaboration and coordination among various stakeholders, including government agencies, the private sector and civil society. The community SOC framework for SMMEs can easily operate within the confines of the NCPF. Having highlighted the need for developing a community SOC, we will now highlight how artificial intelligence can be used to optimise the active cyber defences in the SMMEs' environment.

5.4 Essential ACD framework drivers identified in section 5.

The following essential drivers were identified in section 5 above.

- Lack of automation.
- SOCs are large and complex, necessitating large budgets, advanced skillsets, and advanced technologies.
- Lack of integrated tools, different security tools that cannot be integrated with each other.
- SOC professionals are overwhelmed by too many alerts.
- SOC services are scattered (MSSPs, MDRs), offering different services. For enterprises to fully enjoy the benefits of SOC as a service, they may need to combine the services of MSSPs and MDRs to supplement the security functions of the other resulting in escalating implementation costs.
- For SOC as a service to deliver its services, some organisations might require significant changes in sourcing additional ICT infrastructure.

6. Optimisation of ACD using Artificial Intelligence (AI)

In this section, we will look at how artificial intelligence (AI) can be used to optimise active cyber defence. Firstly, we will briefly discuss artificial intelligence in the context of cybersecurity. Our aim is to highlight how AI technology can be used to optimise active cyber defences in SMMEs' community SOC environment. Digital transformation with themes like digital sales, cloud computing, big data, or artificial intelligence is crucial for the innovative power and future competitiveness of most SMMEs (Lloyd, 2020). According to Burke (2020), SOCs rely on human expertise and specialised skills and knowledge to perform SOC functions. As already mentioned in the previous sections, most organisations and specifically the SMMEs, are unable to benefit from the SOC capabilities due to their limited or lack of advanced cybersecurity skillsets. This is where AI technology fits in in terms of automating complex processes.

AI, as defined by Vähäkainu & Lehto (2019), is the intelligence that is formed artificially to solve complex problems in computers. AI assists in automating complex processes to identify cyberattacks and to react to information system breaches. As the ACD concept aims to increase the automation within an enterprise to help strengthen cyber defences Burke (2020), AI can act as a problem solver in the automation of data processing and threat detection.

According to Burke (2020), the United Kingdom (UK) and the United States of America (USA) are among the countries that incorporate defensive AI technologies into their cybersecurity strategies. Duvenage, Jaquire & von Solms (2020b) state that the convergence of emerging technologies such as AI, the Internet of Things, Big Data, and social media is important in strengthening efforts to cybersecurity defences. Furthermore, Duvenage*, et al*., (2020b) also alludes that these technologies are not only useful for large organisations and nation-states but also for smaller organisations with limited resources.

As the number of cyberattacks grows, so are the costs of discovering the cyberattacks, and emerging technologies may be the answer to finding and reversing this trend. Accenture/Poneman Institute (2019) states that organisations should aim to implement technologies such as AI that reduce the rising costs of discovering a cyberattack. Organisations seek to gain a competitive edge in the market by implementing cyber defence solutions, while cyber threat actors are also actively seeking new ways of launching cyberattacks (Burke, 2020). AI has seen adoption within the cybersecurity domain and can be used for the automated detection and classification of malware and automated threat alerting on network traffic and endpoint data sources (Burke, 2020).

In the following section, we will consolidate the essential drivers that were identified in subsections 3.1, and 5.4.

7. Implications of essential drivers for the development of an active cyber defence framework

7.1 Gaps in the current operating models of SOC as a service

It is worth noting that the essential drivers highlighted in this section are not exhaustive, as the paper is part of an exploratory and ongoing research project. The challenges highlighted in the sections above contribute to the essential drivers necessary towards the building of an active cyber defence framework for SMMEs in developing countries. They will all be consolidated in this section. The essential drivers that will contribute towards the active cyber framework have been consolidated and are listed in Table 1 below. The recommendations will address the identified challenges confronting SMMEs in developing countries. These recommendations will form part of the requirements that can be implemented to develop the ACD community SOC framework.

Table 1: Essential drivers for active cyber defence framework for SMMEs in developing countries.

Essential driver/existing gap	Recommendations
• Different cybersecurity approaches are used to maintain cybersecurity hygiene.	• Develop a unanimous framework for SMMEs in the same industry. This framework will be termed a community SOC framework.
• Lack of financial resources.	• The community SOC framework will relieve the cost burden because the SMMEs will share the cost of implementing the cybersecurity defences.
• Limited or lack of advanced cybersecurity skillsets.	• The advanced skills will be shared across the SMMEs serving a particular community SOC.
• SOC services are scattered (MSSPs, MDRs), offering different services.	• Consolidate SOC services to minimise implementation costs.
• SOC as a service use different tools than those of SMMEs and might not be integrated into the SMMEs'environment.	• Propose a uniform technology for the community SOC. SMMEs in the same industry.
• SOC functions are human-intensive and require huge volumes of data. • Lack of automation, lack of integrated tools, and being overwhelmed by too many alerts. • Complex processes, analysts overwhelmed by too many alerts.	• Automation of data processing and threat detection through AI. Automation will aid in minimising the need for intensive human intervention and reduce the cost of implementation. • Propose essential cyber defence tools that can be integrated into a single platform. • Create universal use cases for the community SOC.

8. Conclusion

Several topics were discussed to achieve the objective of the paper. The following topics were covered: (2) conceptualisation of active cyber defence, (3) cybersecurity landscape of SMMEs in developing countries, (4) the role of SMMEs in developing countries' economies, (5) SOCs as implemented in large enterprises, and (6) optimisation of ACD using AI. The discussed topics, essential drivers and recommendations, viewed in combination, are essentially all factors that decisively influence the development of the ACD community SOC framework.

In concluding this paper, we propose incremental work that will ultimately complete the proposed active cyber defence community framework for SMMEs in developing countries. The proposed future work will detail how the recommendations will be implemented. The paper highlighted and consolidated the essential drivers needed to develop an ACD community framework. Our point of view is that SMMEs need not spend excessive amounts of money to be able to access active cyber defence technologies, nor should a limited budget hinder them from accessing these technologies. The SMMEs should be able to utilise these technologies and still be economically sustainable, hence the proposed solution for ACD community SOC for SMMEs. As far as this paper is concerned, it contributes towards a PhD project of an ACD community framework, that will be expanded on in future work.

References

Abbosh, O. & Bissell, K., 2019. *Securing the digital economy: Reinventing the Internet for Trust.* [Online]
Available at: https://www.accenture.com/_acnmedia/thought-leadership-assets/pdf/accenture-securing-the-digital-economy-reinventing-the-internet-for-trust.pdf
[Accessed 08 01 2023].

Accenture/Poneman Institute, 2019. The Cost of Cybercrime. *Network Security,* 2019(3), pp. 4-4.

Broeders, D., 2021. Private active cyber defense and (international) cyber security—pushing the line?. *Journal of Cybersecurity,* 7(1).

Burke, A., 2020. *Robust artificial intelligence for active cyber defence.* [Online]
Available at: https://www.turing.ac.uk/sites/default/files/2020-08/public_ai_acd_techreport_final.pdf
[Accessed 06 01 2023].

Crowley, C. & Pescatore, J., 2018. *The Definition of SOC-cess? SANS 2018 Security Operations Center Survey.* [Online]
Available at: https://www.sans.org/white-papers/definition-soc-cess-sans-2018-security-operations-center-survey/
[Accessed 27 12 2022].

Crowley, C. & Pescatore, J., 2019. *Common and Best Practices for Security Operations Centers: Results of the 2019 SOC Survey.* [Online]
Available at: https://www.sans.org/media/analyst-program/common-practices-security-operations-centers-results-2019-soc-survey-39060.pdf
[Accessed 10 10 2022].

Denning, D. E. & Strawser, B. J., 2017. Active Cyber Defense APPLYING AIR DEFENSE TO THE CYBER DOMAIN. In: G. Perkovich & A. Levite, eds. *Understanding cyber conflict: 14 analogies.* Washington, DC: Georgetown University Press, pp. 206-208.

Duvenage, P., Jaquire, V. & von Solms, S., 2020a. A Cyber Counterintelligence Matrix for Outsmarting Your Adversaries. *Journal of Information Warfare,* 19(1), pp. 1-10.

Duvenage, P., Jaquire, V. & von Solms, S., 2020b. *Cyber Counterintelligence: Some Contours towards the Academic Research Agenda.* Chester, Academic Conferences and Publishing International Limited.

Harrison, B., 2020. *What is a Security Operations Center (SOC)?.* [Online]
Available at: https://blog.cygilant.com/blog/what-is-a-security-operations-center-soc
[Accessed 06 01 2023].

Herpig, S., Morgus, R. & Sheniak, A., 2020. *Active Cyber Defense- A comparative study on US, Israeli and German approaches.* [Online]
Available at: https://www.kas.de/documents/263458/263507/Active+Cyber+Defense+-+A+comparative+study+on+US,+Israeli+and+German+approaches.pdf
[Accessed 07 12 2022].

Institute of Risk Management, 2014. *Cyber risk.* [Online]
Available at: https://www.theirm.org/media/7237/irm-cyber-risk-resources-for-practitioners.pdf
[Accessed 06 01 2023].

Jaquire, V., Duvenage, P. & von Solms, S., 2021. *Some Cybersecurity Governance Imperatives in Securing the Fourth Industrial.* United Kingdom, Academic Conferences International Limited, pp. 187-194.

Kabanda, S., Tanner, M. & Kent, C., 2018. Exploring SME cybersecurity practices in developing countries. *Journal of Organizational Computing and Electronic Commerce,* 13 07, 28(3), pp. 269-282.

Kaspersky, 2019. *Barriers to an effective Security Operations Center.* [Online]
Available at: https://www.kaspersky.com/enterprise-security/security-operations-center-soc
[Accessed 27 12 2022].

Mihindu, S. & Khosrow-shahi, F., 2020. *Collaborative Visualisation embedded Cost-efficient, Virtualised Cyber Security Operations Centre.* Melbourne, IEEE.

Mimecast, 2022. *Confronting the new wave of cyberattacks: The State of Email Security 2022.* [Online]
Available at: www.mimecast.com
[Accessed 27 12 2022].

Ncubukezi, T., Mwansa, L. & Rocaries, F., 2020. *A Review of the Current Cyber Hygiene in Small and Medium-sized Businesses.* s.l., IEEE Xplore.

Ntingi, N., Duvenage, P., du Toit, J. & von Solms, B., 2022. *Effective Cyber Threat Hunting: Where and how does it fit?.* Chester, Proceedings of the 21st European Conference on Cyber Warfare and Security, pp. 206-213.

Schatz, D., 2014. Thoughts on Managed Security Services Provider Engagement. *ISSA Journal,* pp. 26-31.

StatsSA, 2019. *How large is the small business footprint?.* [Online]
Available at: https://www.statssa.gov.za/?p=12264
[Accessed 11 12 2022].

Vähäkainu, P. & Lehto, M., 2019. *Artificial intelligence in the cyber security environment Artificial intelligence in the cyber security environment.* Stellenbosch, RSA, Proceedings of the 14th International Conference on Cyber Warfare and Security.

Verizon, 2019. *2019 DBIR: Summary of Findings.* [Online]
Available at: https://enterprise.verizon.com/resources/reports/dbir/2019/summary-offindings/
[Accessed 28 12 2022].

Vielberth, M., Bohm, F., Ines, F. & Pernul, G., 2020. *Security Operations Center: A Systematic Study and Open Challenges.* Regensburg, IEEE Access.

Weiss, M. & Muegge, S., 2019. Conceptualizing a New Domain Using Topic Modeling and Concept Mapping: A Case Study of Managed Security Services for Small Businesses. *Technology Innovation Management Review,* 9(8), pp. 55-64.

Participants Prefer Technical Hands-on Cyber Exercises Instead of Organisational and Societal Ones

Jani Päijänen[1], Jarno Salonen[2], Anni Karinsalo[2], Tuomo Sipola[1] and Tero Kokkonen[1]
[1]Institute of Information Technology, Jamk University of Applied Sciences, Jyväskylä, Finland
[2]VTT Technical Research Centre of Finland, Espoo, Finland

jani.paijanen@jamk.fi
jarno.salonen@vtt.fi
anni.karinsalo@vtt.fi
tuomo.sipola@jamk.fi
tero.kokkonen@jamk.fi

Abstract: The current shortage of cybersecurity professionals is about 2 million people worldwide, and in Europe the industry is seeking for about 350 000 skilled professionals. There is also an enormous need for dedicated cybersecurity training courses for existing professionals who wish to acquire completely new skills or maintain their current ones. Due to the lack of new skilled workforce, the current cybersecurity personnel are overworked in their work. In order not to waste the valuable time of cybersecurity professionals with unnecessary training, cyber exercises should be well prepared. This article is based on research conducted in a European collaborative project and more specifically, a cyber exercise organised in early 2022. The purpose of our research was to conduct a preliminary assessment of the participants to learn about their skills and expectations before the cyber exercise. This assessment was used for fine-tuning the exercise. To achieve this, we identified common trends in the participants' interests during the cyber exercise. The preliminary assessment was carried out as a web survey. The responses were cross tabulated to find meaningful indicators related to skills and interests of the participant group. We identified the most and least preferred knowledge areas for both the industry and public sector participants. Our findings show that the most interesting knowledge areas of all respondents were primarily technical in nature (Data Security, Connection Security, System Security), but Organisational Security was also reported. The least interesting knowledge areas were mostly non-technical in nature (Human Security, Organisational Security, Societal Security) but also Component Security was reported. We also enquired about the preferred team size. The majority of the respondents preferred a team size of three to four persons. The preferred single session duration was 46–60 minutes. The results help cybersecurity professionals to match their knowledge needs with the existing cybersecurity proposition and to determine the right and most beneficial training for them. The results also assist the providers of cyber training and other exercises to describe the targeted development of specific cybersecurity and other knowhow in a coherent, standard-like, way.

Keywords: Cybersecurity, Cyber Range, Cyber Exercise, Capture-The-Flag, Skills

1. Introduction

One of the most valuable assets in cybersecurity is the knowhow of the people. In a workplace, this naturally relates to the staff members. In addition to the global lack of competent cyber security experts (Blažič, 2021), there is a distinct gap between the appropriate skills of the cyber security professionals and the cyber security requirements of the occupations they hold (Nurse et al., 2021). For this, one of the goals in the Cyber Security for Europe (CyberSec4Europe) project was to specify learning objectives and competences required to develop and enhance cybersecurity skills for different profiles and roles. Considering the development and enhancing of skills of the workforce, there is a need not only to improve academic education but also the effectivity of hands-on courses. Cybersecurity exercises are one of the most effective modes for training the required skills and knowhow in the cybersecurity domain, especially under a stressful cyber-attack situation. Cybersecurity exercises arranged in cyber ranges enable organisations to be trained for reacting in situations when their assets or the whole organisation are experiencing cybersecurity incidents.

The Flagship 1 exercise, organised in January 2021 in connection with the CyberSec4Europe project, was an online cyber exercise. It featured the technical federation of cyber ranges so that participants in multiple locations across Europe could use the environment, which gathered positive feedback from the short after-survey (Kokkonen et al., 2023). The exercise this article focuses on, Flagship 2, was organised in January 2022. It focused on protecting the operations of a fictional train operator by offering (i) a cybersecurity exercise and (ii) an open external cyber security analyst activity using Capture-the-Flag (CTF) concepts (Kokkonen et al., 2022).

Both cybersecurity exercises and CTF challenges are useful tools in teaching cybersecurity topics. Firstly, the organisational aspects of cyber exercises have been studied in addition to the more traditional technical ones. Karjalainen et al. (2022) emphasise the evolution of cybersecurity exercises into organisational learning experiences. They conducted a survey on learning during an on-line exercise aimed at an organisation's various

work roles. Furthermore, Karjalainen and Kokkonen (2020) have proposed cyber arenas as learning environments to simulate organisations' environments to further their learning goals. According to them, this expansive use of cyber exercises provides "core knowhow elements of the cyber environment" for successful organisational exercises. Secondly, Capture-the-Flag (CTF) exercises are a well-known method of teaching cybersecurity in a class environment. The participants are presented with challenges related to cybersecurity, usually in a competitive setting (Mirkovic and Peterson, 2014). Using CTFs in cyber ranges for risk estimation shows that the combination is useful in many applications (Di Tizio et al., 2020). However, it seems that non-technical aspects should be included so that all the stakeholders could take part in exercises (Švábenský et al., 2021).

The importance of measuring learning should be highlighted when developing cyber exercises. Web questionnaires before and after a cyber exercise have been identified as useful tools to measure the experience of the exercise participants (Karjalainen et al., 2020). Using self-assessment questionnaires before an exercise is useful when setting learning objectives for an exercise (Vykopal et al., 2017). Their usage has been demonstrated to be a useful way to gather information before and after an exercise to evaluate learning outcomes (Vielberth et al., 2021).

The immediate feedback of Flagship 2 has been analysed by Kokkonen et al. (2022). The questionnaire was short and aimed to map out the participants' experiences and opinions about the usefulness of the exercise. The main finding is that the participants indicated that they obtained new skills.

In this article, we describe and analyse the results of the implemented Flagship 2 cyber exercise preliminary survey, which was designed by Karinsalo et al. (2022). The objective was to collect information from the registered participants prior to the exercise. This information consisting of more detailed information about the participants' background, expertise and interest areas could then be used to customise the exercise to better match the expectations and interests of the participants. The scientific contribution of this article comprises the results from the survey and the first analysis with the objective to provide meaningful activities for future cyber exercises through customisation. We selected the following research questions to guide our work:

1. How can we improve the quality of a cyber-exercise through pre-assessment of the participants?
   - How can we enhance the participant experience within a specific cyber-exercise?
2. What are the most and least interesting topics in a cyber-exercise?
3. How do the cyber-exercise needs of academia and research differ from the needs of industry?

The article is structured as follows. We begin the second chapter by describing the methodology. This includes the cyber exercise selected as the target for our work as well as the questionnaire generation, survey invitation and the analysis processes of the initial results. The third chapter focuses on the survey results and the analysis of data, providing the answers to our research questions. Finally, in chapter four we discuss the most significant findings from the survey and potential differences of our results when compared to previous research as well as ideas for future research before concluding the article in chapter five.

2. Methods

Flagship 2 consisted of two parallel activities. The primary activity was the cyber exercise that used a cyber range, a prepared complex learning environment. The exercise participants were placed into teams and coached by a dedicated team coach, who guided the team with questions. The exercise was targeted at CyberSec4Europe affiliates. The secondary activity was a capture-the-flag (CTF) exercise, where the participants were offered six digital samples or evidence exported from the exercise environment. The CTF activity was open for everyone. The participants of the CTF activity worked independently analysing the samples in a virtual machine with the focus on digital forensics. Later in this article, we refer to the CTF participants as *analysts*. Both activities required registration and lasted for two consecutive days. The participants were not scored nor otherwise assessed to avoid unnecessary competition, and thus a safe and comfortable environment for learning was created.

In order to define the cybersecurity skill levels of the participants prior to the exercise, as well as determine their expectations for it, we conducted a survey with six background questions and six research questions. We have described the survey and process thoroughly in Karinsalo et al. (2022), but the following paragraphs will summarise the survey content and process briefly.

The first part of the survey consisted of background questions such as participant email address (question #1), which was needed to identify the participants for after-survey purposes (not covered in this article). This was followed by background questions consisting of multiple-choice answers and covering the participants'

educational background (question #2), primary sector of the participant organisation (question #3), knowledge (question #4) and technical skill levels (question #5) related to cybersecurity exercises/hackathons, and the primary job role of the participant (question #6).

The second part of the survey introduced six research questions related to the cybersecurity exercise. This part covered the primary knowledge area development/improvement that the participant was interested in (question #7), preferred role within the exercise (question #8), ideal number of participants in the exercise teams (question #9) and preferred average length of the exercise sessions (question #10). The last two questions inquired about the knowledge area that the participant was least interested in (question #11), and there was an open question for any thoughts that the participant might have regarding the exercise (question #12).

We conducted the survey for the live exercise participant group during the timespan of January 11 – 28, 2023, and an identical survey for the analyst group during the timespan of January 19 – 24. The exercise hosts sent survey links in an invitation email sent to registered participants of both participant groups of the cyber exercise, and both surveys were conducted using the Questback online survey service (available at https://www.questback.com/). None of the survey questions were mandatory: respondents answered to as many questions as they chose.

The analysis was carried out by cross tabulation using the Microsoft Excel spreadsheet program. The answers to the questions were compared within and between categories. The respondents were divided into three categories. The category “Industry” included participants from companies in multiple domains while “Public sector” included participants from education, research, government, church, and non-profit organisations. The third category was used for respondents who did not disclose their organisation.

3. Results

3.1 Results from the survey process

The total number of surveyed participants was 82, consisting of live exercise participants and analysts. We received a total of 55 responses, yielding a response rate of 67%. Since the survey form, survey process and invitations were identical in both target groups, we have merged both groups into one respondent group. Two out of 55 respondents chose not to disclose their educational background, and nine respondents chose not to disclose their business sectors. Consequently, the analysis below excludes these respondents, resulting in 53 in the analysis concerning exercise participant education, and 46 in the analysis of their respective business sectors. However, we have included the non-disclosed as an independent group when calculating, e.g., responses by business sector.

3.2 Background data from the survey

The first background question covered the educational background of the participants. The most common educational background in total was Master’s Degree (EQF7) with 20 respondents (37.7% out of 53), Bachelor’s Degree (EQF6) with 19 respondents (35.8%) was quite close to the previous, and the third most common education was Doctoral Degree (EQF8) with 9 respondents (17%), Vocational education was fourth with 4 respondents (7.5%), and one respondent specified “Upper secondary school” in the open text field as their educational background (2%), which is quite similar to one of the available options, namely Vocational education (EQF4).

The second background question covered the business sector of the participants. The survey response data was minimised to include three business sector options: Industry, Public Sector, and an option to not to disclose the information. Industry was reported by 25 respondents (54% out of 46), and 21 respondents (46%) reported public sector as their home organisation’s business sector. The split between exercise and CTF participants by business sector is shown in Table 1.

Table 1: Responses split by Business sector between analysts and exercise participants.

Business Sector	Analysts	Exercise	Total
Industry	16	9	25
Public Sector	5	16	21
Not Disclosed	6	3	9

Business Sector	Analysts	Exercise	Total
Total	**27**	**28**	**55**

The participants' knowledge level, including the understanding of concepts and types in cybersecurity exercises and hackathons was enquired in the third background question. The most common selected option was Intermediate (Apply/Analyse) with 26 respondents (47.2%, when n=55), Entry level (Remember/Understand) with 25 (45.5%), and Expert (Evaluate/Create) with four respondents (7.2%).

The fourth background question covered respondents' self-opinion of their technical skill level (e.g., usage of operating systems and IT environments, etc.) regarding cybersecurity exercises and hackathons (Table 2). The revised Bloom's taxonomy (Anderson et al., 2001) was used when categorising questions three and four.

Table 2: Technical skill level by analysts and exercise participants

Technical skills level	Analyst	Exercise	Total
Entry level (Remember/Understand)	7	11	18
Intermediate (Apply/Analyse)	18	13	31
Expert (Evaluate/Create)	2	4	6
Total	**27**	**28**	**55**

The fifth background question covered the participants' job role. The single-choice options followed the NIST - National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework (Newhouse et al., 2017) categorisation. Two respondents reported a job role outside of the NICE framework (Table 3). In total 11 respondents did not disclose their job role.

Table 3: Job roles reported by analysts and exercise participants

Job role	Analyst	Exercise	Total
Analyse (AN)	1	6	7
Investigate (IN)	0	4	4
Operate and Maintain (OM)	9	3	12
Oversee and Govern (OV)	6	5	11
Protect and Defend (PR)	2	3	5
Securely Provision (SP)	2	1	3
Communications	0	1	1
Teaching and research	0	1	1
Not disclosed	7	4	11
Total	**27**	**28**	**55**

3.3 Research results from the survey

The research survey questions topics were (i) most interesting knowledge area to improve or develop, (ii) role to proceed in the interesting knowledge areas, (iii) ideal number of participants for the exercise teams, (iv) ideal duration of the exercise situation before the situation develops, and (v) least interesting knowledge area to improve or to develop. These are presented in the next paragraphs.

3.3.1 Most interesting knowledge areas

Survey question #7 was multiple-choice, where the respondents were guided to select from one to three most interesting knowledge area options that the respondent was most interested in developing or improving during the exercise. The respondents were informed that the goals, and thus the contents of the Flagship exercise and CTF events were already set.

The nine knowledge area options offered in static order to the respondents have been presented in Table 4. Table columns are Public Sector (PS), Industry (I), and "I prefer not to disclose this information" (ND).

In total System Security was voted 36, Data Security 29, and Connection Security 23 times. Public sector and industry respondents voted System Security 16, and Data Security 14 times. Both Human Security and Organisational Security were voted eight times. Analyst respondents voted System Security 15 times. Connection and Data security were both voted 11 times.

Table 4: Number of selected most interested knowledge area options

Knowledge area	Exercise			Analyst			Total
	PS	I	ND	PS	I	ND	
Data Security	9	5	1	1	10	3	29
Software Security	5	2	2	3	6	1	19
Component Security	0	2	1	0	0	0	3
Connection Security	6	2	0	4	7	4	23
System Security	9	7	1	4	11	4	36
Human Security	5	3	1	0	3	3	15
Organisational Security	4	4	1	2	5	2	18
Societal Security	3	1	1	0	1	1	7
Operate and maintain	2	0	1	1	1	0	5
Total	**43**	**26**	**9**	**15**	**44**	**18**	**155**

3.3.2 The role to develop in Flagship 2

The single choice question #8 surveyed the primary role that the respondents wanted to develop, by choosing the most interesting knowledge area options. The question followed the Newhouse et al., (2017) categorisation. Exercise participants selected Securely Provision (SP) nine (32.1%, n=28) times, and Investigate (IN) six (21.4%) times. Analysts selected seven times (25.9%, n=27) both Operate and Maintain (OM), and Securely Provision (SP).

Table 5: Role to develop by selected knowledge areas

Role to develop	Exercise			Analyst			Total
	PS	I	ND	PS	I	ND	
Analyse (AN)	3	2	1	1	3	1	11
Investigate (IN)	2	1	1	0	5	2	11
Operate and Maintain (OM)	2	1	0	3	0	0	6
Oversee and Govern (OV)	4	1	0	0	2	1	8
Protect and Defend (PR)	4	4	1	0	5	2	16
Securely Provision (SP)	0	0	0	1	0	0	1
Collect and Operate (CO)	1	0	0	0	1	0	2
Total	**16**	**9**	**3**	**5**	**16**	**6**	**55**

3.3.3 Ideal number of participants in an exercise team

In question #9 we asked about the ideal number of participants for the exercise teams (Table 6). This was a single choice question. Exercise participants selected the option “3-4 participants” 19 (67.9%, n=28) times, and “more than six” seven (25.0%) times. Analysts selected the option “3-4 participants” 20 (74.0%, n=27) times.

Table 6: Ideal number of participants in an exercise team

Number of participants	Exercise			Analyst			Total
	PS	I	ND	PS	I	ND	
1–2	0	1	1	0	1	1	4
3–4	12	5	2	4	12	4	39
5–6	4	3	0	0	1	1	9
more than 6	0	0	0	1	2	0	3
Total	**16**	**9**	**3**	**5**	**16**	**6**	**55**

3.3.4 Exercise session development

Cyber security exercises may have phases or sessions that develop either behind the scenes or in a way that is visible to the participant. In the single choice question #10 we asked the preferred session duration (Table 7). Exercise respondents selected the option "46-60 minutes" 11 (39.2%, n=28), and option "31-45 minutes" 10 (35.7%) times. Analysts reported "46-60 minutes" 13 (48.1%, n=27) and option "61-90 minutes" seven (25.9%) times.

Table 7: Exercise session duration

Session duration	Exercise			Analyst			Total
	PS	I	ND	PS	I	ND	
16-30 minutes	2	0	0	1	1	0	4
31-45 minutes	5	4	1	0	3	1	14
46-60 minutes	6	4	1	3	9	1	24
61-90 minutes	2	1	1	0	3	4	11
more than 90 minutes	1	0	0	1	0	0	2
Total	**16**	**9**	**3**	**5**	**16**	**6**	**55**

3.3.5 Least interesting knowledge area

Question #11 was similar to the question #7 but asking the opposite preference, i.e. the least preferred knowledge area development or improvement options. The respondents were guided to select from one to three options.

The exercise participants voted both Human Security and Societal Security nine times (Table 8). Component security was voted eight times. Data Security, and System Security, and Operate and maintain were each voted seven times. Analysts voted Societal Security 11, Organisational Security 10 times. Both Human Security, and Component Security were voted seven times.

Table 8: Number of selected least interesting knowledge area options

Knowledge area	Exercise			Analyst			Total
	PS	I	ND	PS	I	ND	
Data Security	4	3	0	1	2	1	11
Software Security	2	0	0	0	1	1	4
Component Security	4	2	2	1	4	2	15
Connection Security	2	1	1	0	3	1	8
System Security	3	3	1	0	0	0	7
Human Security	5	4	0	2	4	1	16
Organisational Security	2	2	0	1	6	3	14
Societal Security	5	4	0	4	6	1	20
Operate and maintain	5	2	0	0	3	2	12
Total	**32**	**21**	**4**	**9**	**29**	**12**	**107**

3.3.6 Open Question

Question #12 was an open question to enquiry the respondents' thoughts about the forthcoming event. We received 11 responses. These responses are not covered in this article.

3.4 Findings

Our findings show that the most interesting knowledge areas reported by all respondents were primarily technical in nature (Data Security, Connection Security, System Security), but Organisational Security was also reported. The least interesting knowledge areas were mostly non-technical in nature (Human Security, Organisational Security, Societal Security) but included also Component Security.

Key data from the public sector or industry respondents are shown below.

Exercise participants ($n=25$):

- Ideal number for participants for the exercise teams = 3-4 (68.0%)
- Average exercise session (or phase) duration = 46-60 minutes (44.0%) or 31-45 minutes (40.0%)
- Most interesting knowledge development area = System security (16 votes), Data security (14 votes)
- Least interesting knowledge development area = Societal (9 votes) and Human security (9 votes).

Analysts ($n=21$):

- Ideal number for participants for the exercise teams = 3-4 (76.1%)
- Average exercise session (or phase) duration = 46-60 minutes (61.9%) or 61-90 minutes (33.3%)
- Most interesting knowledge development area = System (15 votes) and Connection (11) and Data security (11)
- Least interesting knowledge development area = Societal security (10 votes), Organisational security seven, and Human security six votes

Among industry partners, the least interesting topic was societal security. It covers cybercrime, law, policy, and privacy topics. Understanding of societal security is needed when establishing organisation's (cyber) security controls, procedures, and guidelines, and when responding to changing legislation requirements. The requirements are transformed to company's internal (cyber) security or privacy requirements, which are then handled and implemented in the organisation's way.

Among industry partners, human security and organisational security are the most diverging. Human security includes, e.g., identity management, social engineering, awareness and understanding, and usable security and privacy. What comes to end-users, awareness and understanding are key concepts. They are hoped e.g., not-to-fall into phishing emails or CEO frauds. Social engineering has been, even lately, successfully used as the mechanism to attack organisations. When such vulnerabilities in human security are exploited, then organisational security functions will activate, e.g., in the form of business continuity and incident management.

4. Discussion and future research

The information gained from this survey shows that the use of pre-assessment questionnaires in cybersecurity exercises can be a valuable source of information about the interests and needs of the participants. Especially with remote participation, a questionnaire can provide deeper understanding of their needs, and this way direct the development of future exercises with new requirements. Such requirements include areas of interests and pacing of the exercise.

The preference of technical hands-on exercises could be explained by the profile of the respondents. Moreover, the use of cyber exercises to improve organisational readiness could be a distant and unfamiliar topic to technically minded participants. Consequently, we should not draw the conclusion that organisational exercises are useless or unwanted in general. Based on our experience, even the most diverging elements, societal security and human security can be offered to specialists, managers, and directors in an exercise. Participants' reflection sessions after an exercise may raise the question whether the participant's organisation could detect or respond to the cyberattacks similarly or more efficiently than in the exercise, which could eventually improve the organisation's (cyber) security.

Both industry and public sector respondents found societal security the least interesting topic. This finding contrasts with Budde et al. (2023), who stated that societal values are interesting especially among industry

respondents in the cybersecurity context. This is most likely a result from different survey target groups, which in our case consist of people that had already declared their interest in a cyber exercise by completing the registration process, while the survey by Budde et al. was targeted to a broader group of cybersecurity experts in Europe. Since Budde et al. did not report detailed information about the respondent profiles, we can assume that their target group, more than ours, may have represented people from management-level personnel and thus been oriented to wider aspects of cybersecurity than ours.

As a conclusion, it could be that preconceptions of CTF events causes the technical aspect being seen as the most important part of the exercise.

The order of choices in the questionnaire could cause bias towards the first options. With larger numbers of respondents, randomized order of questions might provide a more balanced set of answers.

As the above discussion shows, there are still open questions in the motivation of cyber exercise participants. Further verification of the difference in interests in technical and societal knowledge is needed. Future research could also include an analysis of questionnaires during and especially after an exercise.

5. Conclusion

Our results display that a pre-assessment can (help to) improve the quality of a cyber exercise. The organisers will be able to profile the participants before the exercise and adjust the profiles to be as useful as possible to them. By asking about the expectations of the participants beforehand, the organisers can identify potential challenges. Knowing the participants' professional experience can also facilitate matching the skill level of the tasks to their skills.

The participant experience can be enhanced by creating tasks that fit the participants' profiles. The exercise roles can be distributed among the participants more effectively so that exercise participants and analyst positions are filled with suitable candidates. The preference for group size can also be considered, but the organisers should also use their own experience for optimal learning outcomes.

This study identified the three most interesting topics in a cyber exercise: System Security, Data Security and Connection Security. The least interesting topics were Societal Security, Human Security, Organisational Security and Component Security, the last two getting the same number of votes. When inspecting the differences between the Industry and Public Sector categories, both preferred System Security, Data Security, Connection Security. However, there were some differences between the categories: Industry also included Organisational Security in the top selection, while Public Sector contained also Software Security as a preference.

There were similarities between the Industry and Public Sector categories. Both listed Societal Security as the least preferred topic. In addition, Human Security was among the three least favourites. Curiously, the Industry respondents voted Organisational Security to be among both the most and the least preferred topics. Such polarisation could originate from the diverse work roles that the participants hold in their organisations. In the future, it could be beneficial to ask for just one top preference from the participants to reduce noise and to obtain hard preferences.

Acknowledgement

This research was partially supported by the Cyber Security Network of Competence Centres for Europe (CyberSec4Europe) project funded by the European Union under the Horizon 2020 SU-ICT-03-2018 Programme Grant Agreement No. 830929.

The authors would like to thank Ms. Tuula Kotikoski for proofreading the manuscript.

References

Anderson, L., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. R., Pintrich, P. R., Raths, J. and Wittrock, M. C. (eds.) (2001) A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives, abridged edition, Addison Wesley Longman, New York.

Budde, C., Karinsalo, A., Vidor, S., Salonen, J. and Massacci, F. (2022) "Consolidating cybersecurity in Europe – A case study on job profiles assessment", *Computers & Security*. https://doi.org/10.1016/j.cose.2022.103082

Di Tizio, G., Massacci, F., Allodi, L., Dashevskyi, S. and Mirkovic, J. (2020) "An experimental approach for estimating cyber risk: a proposal building upon cyber ranges and capture the flags", *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pp. 56–65. https://doi.org/10.1109/EuroSPW51379.2020.00016

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

Karinsalo, A., Saharinen, K., Päijänen, J. and Salonen, J. (2022) "Pedagogical and self-reflecting approach to improving the learning within a cyber exercise", *Proceedings of the 21st European Conference on Cyber Warfare and Security, ECCWS 2022*. Vol. 21, No. 1, pp. 105–114. https://doi.org/10.34190/eccws.21.1.221

Karjalainen, M., Puuska, M. and Kokkonen, T. (2020) "Measuring Learning in a Cyber Security Exercise", *2020 12th International Conference on Education Technology and Computers (ICETC'20)*, pp. 205–209. https://doi.org/10.1145/3436756.3437046

Karjalainen, M. and Kokkonen, T. (2020). "Comprehensive Cyber Arena; the Next Generation Cyber Range", *Proceedings of the 5th IEEE European Symposium on Security Privacy Workshops, Virtual Event, 7–11 September 2020*. https://doi.org/10.1109/EuroSPW51379.2020.00011

Karjalainen, M., Kokkonen, T. and Taari, N. (2022). "Key Elements of On-Line Cyber Security Exercise and Survey of Learning During the On-Line Cyber Security Exercise", Lehto, M. and Neittaanmäki, P. (eds.) *Cyber Security: Computational Methods in Applied Sciences.* Springer, Cham. https://doi.org/10.1007/978-3-030-91293-2_2

Kokkonen, T., Päijänen, J. and Sipola, T. (2022) "Multi-National Cyber Security Exercise, Case Flagship 2", *2022 14th International Conference on Education Technology and Computers (ICETC 2022)*. Forthcoming 2022. https://doi.org/10.1145/3572549.3572596

Kokkonen, T., Sipola, T., Päijänen, J. and Piispanen, J. (2023) "Cyber Range Technical Federation: Case Flagship 1 Exercise", Dimitrakos, T., Lopez, J., Martinelli, F. (eds.) *Collaborative Approaches for Cyber Security in Cyber-Physical Systems*. Advanced Sciences and Technologies for Security Applications. Springer, Cham, pp. 1–13. https://doi.org/10.1007/978-3-031-16088-2_1

Mirkovic, J. and Peterson, P. (2014) "Class Capture-the-Flag Exercises", *2014 USENIX Summit on Gaming, Games, and Gamification in Security Education (3GSE '14)*.

Newhouse, W., Keith, S., Scribner, B. and Witte, G. (2017) *National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework*. NIST SP 800-181. Gaithersburg, MD: National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-181

https://doi.org/10.6028/NIST.SP.800-181

Nurse, J., Adamos, K., Grammatopoulos, A. and Di Franco, F. (2021) "Addressing the EU cybersecurity skills shortage and gap through higher education", European Union Agency for Cybersecurity (ENISA), https://www.enisa.europa.eu/publications/addressing-skills-shortage-and-gap-through-higher-education

Švábenský, V., Čeleda, P., Vykopal, J. and Brišáková, S. (2021) "Cybersecurity Knowledge and Skills Taught in Capture the Flag Challenges", *Computers & Security*, Vol. 102, Art. 102154. https://doi.org/10.1016/j.cose.2020.102154

Vielberth, M., Glas, M., Dietz, M., Karagiannis, S., Magkos, E. and Pernul, G. (2021) "A Digital Twin-Based Cyber Range for SOC Analysts", *Data and Applications Security and Privacy XXXV. DBSec 2021.* Lecture Notes in Computer Science. Vol. 12840. Springer, Cham, pp. 293–311. https://doi.org/10.1007/978-3-030-81242-3_17

Vykopal, J., Vizvary, M., Oslejsek, R., Celeda, P. and Tovarnak, D. (2017) "Lessons learned from complex hands-on defence exercises in a cyber range", *2017 IEEE Frontiers in Education Conference (FIE)*. https://doi.org/10.1109/FIE.2017.8190713

A Reflection on Typology and Verification Flaws in Consideration of Biocybersecurity/Cyberbiosecurity: Just Another Gap in the Wall

Lucas Potter[1], **Kimberly Mossburg**[1-2] **and Xavier-Lewis Palmer**[1-2]
[1]BiosView, Oswego, USA
[2]Cyber Solutions Academy, Franklin, USA

lpott005@odu.edu
kmorring@gmail.com
xpalm001@odu.edu

Abstract: Verification is central to any process in a functional and enduring cyber-secure organization. This verification is how the validity or accuracy of a state of being is assessed (Schlick, 1936; Balci, 1998). Conversely, breakdown in verification procedures is core to the interruption of normal operations for an organization. A key problem for organizations that utilize biology as an interlock within their systems is that personnel lack sufficient ability to verify all practically relevant biological information for procedures such as a nurse logging a blood draw, or a molecular biology technician preparing agar to culture microbes for study. This has several implications, one of which is our diminished ability to approximate and defend against emerging biologically-linked cyberthreats. These could be in the form of mis- or dis-information, contaminants, or calculated threats to vital supplies. Two important questions to ask are: "What may be the implications of diminished ability to undergo strict verification measures (such as triple redundancy and technological distancing)." And "how does this impact our ability to anticipate and make changes for verification of biological processes?" This paper aims to discuss key areas where verification gaps exist and how to bridgethos gaps. Towards this, we cover data integrity, implications of the lack of verification, triple redundancy, technological distancing, biosafety concerns, and more. All of this will factor into the ability of organizations with proximity to biosecurity to anticipate national changes to biological processes that are nationally relevant.

Keywords: Verification, Typology, Biocybersecurity, Cyberbiosecurity, Vulnerability, Ontology

1. Introduction

There exists a key problem within biosafety among corporate and national officials, and in the general population (including community biolab members) in which many do not fully understand or appreciate the ability to verify biological information. This leaves us with gaps to address in the field on multiple fronts. Namely: in industry, community bio labs, and government. This has several implications, including a diminished ability to approximate and defend against emergent threats, be they mis or disinformation, or contaminants. An important question to ask is what the implications of the lack of ability may be to undergo strict measures such as triple redundancy and technological distancing. Further, how does this impact our ability to anticipate and make national changes for biological processes?

Previous papers have explored the idea of the future of threats at the nexus of both biological and cyber domains (Bevilacqua et al, 2022; Duncan et al, 2019; Griffin, 2023; Palmer, 2021; Peccoud et al, 2018; Reed et al, 2019; Richardson et al, 2019). Now is a relevant moment to explore how that nexus can be bridged with other avenues of aggression, and how those can be effectively defended against. Through the polymorphisms that biology presents in data processing, storage, and presentation, these potential avenues increase appreciably as prior demonstrations and advancements within the last 10 years have demonstrated, as stated by Murch (2018). Potter & Palmer demonstrated that the early threats can be approximated, and wargaming has been shown to be extendable to biocybersecurity (2021).

In cybersecurity and its offshoot, cyberbiosecurity, typology refers to the classification and categorization of different types of malicious cyber activities that can pose a risk to the security of biological systems and data. This includes identifying the different types of cyber threats that can target these systems, such as malware, phishing, and ransomware, as well as the tactics, techniques and procedures used by attackers. Understanding the different types of cyber threats that can target biological systems is important for several reasons. It allows organizations to better identify and defend against specific types of cyber threats that can target these specific types of data. It helps organizations prioritize their cyberbiosecurity efforts by focusing on the types of cyber threats that are most likely to occur and have the greatest potential impact. It allows for the development of more effective cyberbiosecurity tools and strategies by providing a better understanding of the tactics, techniques, and procedures used by attackers. It can help in detecting and responding to cyber incidents that may compromise biological systems and data. It also allows organizations to stay up-to-date on the latest trends and innovations in the field of cyberbiosecurity.

In biocybersecurity typology can be used to classify different types of biological threats such as viruses, bacteria, and toxins, as well as different types of cyber threats such as malware and phishing attacks. This can aid in identifying and understanding the characteristics of these threats, and in developing appropriate countermeasures to protect against them. Additionally, typology can be used to classify different types of organisms, such as bacteria and fungi, which can be used in biotechnology applications such as the production of biofuels and other products. This can help in the development of security measures to protect against the unauthorized release or misuse of these organisms. However, in this new hybrid field of biocybersecurity typology and verification may prove to be problematic. There is no systemic method to implement the policy, standards of practice, or legal considerations for protecting the bioeconomy. It is arguable that even with the newest developments in Artificial Intelligence (AI) there may never be. Experts in the field are generally not aware of each other's expertise, perspectives and even mutually supported opportunities to work together to provide positive outcomes. It is imperative to create standards for cyberbiosecurity and establish a structured discipline that can be supported by various individuals.

In summary, typology in cyberbiosecurity (CBS) and biocybersecurity (BCS) represents ways to categorize, analyze and understand the different types of threats that can target biological systems and cyber-linked data, which can aid in better defense and incident response strategies. The paper includes a literature review concerning the typology of cybersecurity in general. The literature review delves into the incentivization of cybersecurity, as well as network considerations and the impact of human behavior on cybersecurity. This review is to be used as a means to explore how analogues of cybersecurity may exist in biocybersecurity. The review then explores why one should care about the possibility for people to exploit these openings in security. It should be noted that because BCS/CBS is a new field there are extremely limited resources available to discuss its impact in the field of cybersecurity as a whole.

This paper intended to be used as a stepping stone to begin that discussion. Considering the burgeoning expanse of the multi-trillion dollar global bioeconomy(Kacprzak et al, 2022), it is entirely likely that most typological security systems will one day be linked within the cyber-bio nexus. To dutifully cover those issues, the most pertinent subdomains and examples have been selected. Some may seem only tangentially related to BCS, but they are fields that should soon face integration with biologically relevant endeavors. We begin with a literature review on cybersecurity that discusses how a cybersecurity typology can refer to many things, but that it is only complete when that typology can be understood by all parties. The paper then moves to the incentivization of cybersecurity which delves into the incentives (economic or otherwise) that organizations look to when determining the use of cybersecurity within their organization. This is followed by a review of network considerations of security typology, which breaks down different network types and components and how they are impacted by cybersecurity typology. The final topic that is covered in the literature review is that people and their behaviors are the most common threat to keeping data secure. We follow this with a discussion of potential limitations.

2. Typology Literature Review

Within cybersecurity there are many different typologies that can be considered. In the scoping review conducted here, it was discovered that no two authors have the same point of view regarding what the typology for cybersecurity should entail.

Curwell (2022) says that a typology refers to how a system can be broken down into different parts. Current research indicates that cybersecurity typology could refer to many different things including: the governances of cybersecurity, the threats and risks that are associated with cybersecurity, economic impacts of cybersecurity, reasons why organizations do (or do not) invest in cybersecurity, different types of networks and how they are impacted by cybersecurity, how cybersecurity resources are managed within an organization, how cybersecurity policies impact parts of the network (specifically the DMZ and firewalls), and managing the impact of humans on cybersecurity. In this scoping review, the authors found that there is no agreement as to how cybersecurity should be broken up and the best ways to implement cybersecurity policies effectively in an organization's environment.

Curwell (2022) states that a typology is only complete when it can be easily understood by all parties and can then be converted into a decision-making process without needing much additional clarification. This is an important note, as many cybersecurity implementations in both the national and corporate domains require substantial interpretation in order to be implemented at all, much less at an effective scale. Typologies should be written to be understood by the intended audience, whether that be data scientists, forensic investigators,

CISOs, CEOs, or cyber security analysts, and then the information should be able to be addressed on a case-by-case basis. AI is one way to automatically break some of these rules down- however this runs into substantial issues with traceability of the interpreted requirements. Curwell (2022) argues that the problem with creating a singular cybersecurity typology is that they will never be complete since there are constant changes occurring in how threat actors are working to avoid detection, as well as how the technology changes over time. Additionally, if a particular threat is not going to actually pose a risk to an organization, then that company does not require a typology for that particular cybersecurity threat. One of the primary issues with typologies in the fields of BCS and CBS is that the threats are not typically connected between cyber connections and biological materials. Since there is such a vast number of cyber threats that can impact individual organizations, it is near impossible to find one set typology that can apply across the scope of the entire field of cybersecurity.

One of the areas within cybersecurity where a typology is present is in the different governance models. Eggenshwiler (2018) states that there are three governances that are often seen in cybersecurity: hierarchical, market and multi-stakeholder. A hierarchical governance is one where there is top-down regulation (as in with specific rules, policies or laws). It is an authoritative system of command and control with strict accountability. This could be seen in military establishments, regional departments, and important in incidents that require collaboration between incidents that involve governments, private industry, and nonprofit organizations. Government policy, including international treaties (like GDPR which was adopted in 2016), on the use of cybersecurity are examples of hierarchical governance. In contrast, market governance is bottom-up regulation. Eggenschwiler (2018) argues that this accounts for decentralization and the creation of independent autonomous units. Malware defense and packet routing are some areas where market governance is appropriate. Multi-stakeholder governance is where the majority of cybersecurity governance lies. It is a combination of hierarchical and market governance and is based on an exchange relationship between stakeholders as a way to find solutions to problems. These interactions are based on trust, reciprocity, and consensus-based decisions. The management of DNS (Domain Name System) is an example of multi-stakeholder governance.

3. Incentivization of Security Typology

Further research indicates that some people believe that cybersecurity can be broken down into six separate areas that determine why organizations do (or do not) invest in cybersecurity for their company. Wessels et. al. (2021) state that the areas that cybersecurity can be broken down into are: economic, normative, historic, feasibility, network externalities, and the idea that there are competing issues & solutions within cybersecurity. The idea of topology was applied in this instance to try and analyze why cybersecurity investments are, or are not, made within an organization. Wessels, et. al. (2021) argued that the feasibility of the implementation of different security measures was a very impactful point, as organizations made estimations about proposed interventions and how they could be incorporated into the business' current infrastructure. One critical point to note is that an organization must be aware of both the cybersecurity problems and solutions that are feasible within that organization, because without that awareness it is near impossible to determine why cybersecurity is a worthwhile investment.

Wessels, et. al. (2021) also stated that there are a wide range of incentives for the use of cybersecurity within an organization and each stakeholder will look at these inceptives differently. One of the biggest incentives deals with the global economy. Economic incentives allow for the maximum profit in both reputation and operations for an organization when deciding how to invest in cybersecurity. These incentives look for the impact of cybersecurity on costs, reputation, clients and any competitive advantage that might be utilized. According to Cremer, et. al. (2022) cybercrime had a cost of almost 950 billion USD in 2020 to the global economy and also caused a rise in cybersecurity insurance. In order to better avoid negative economic impact, it is critical to have a robust cybersecurity defense system at corporate, national and global levels.

Unfortunately, there is limited data available on cyber risks to the international economy. Cremer, et. al. (2022) argued that cybersecurity is still an emerging risk and does not have a lot of historical data. When a breach occurs, institutions do not necessarily publish the incident - creating a lack of data which proves to be a challenge for research, risk management and cybersecurity. This could be an issue when it comes to the development of future cybersecurity methods or policies, as the threats to be defended against are not public knowledge until they are an onerous enough problem to cause widespread disruption. There are so many different cyber-attacks and data breaches that can occur, ranging from ransomware to DDoS attacks, that any data to help prevent future economic loss to organizations would help guide these policies that in part assure the stability of the global economy. Mandatory reporting of the details of cyber incidents could help improve understanding,

awareness and loss prevention as well as allowing cyber risks to be better understood (which in turn enables better research).

From a business standpoint, there is an investment challenge when it comes to cybersecurity. Fernandez De Arroyabe, et. al. (2023) states that organizations must determine how much of an investment in cybersecurity is needed to provide sufficient protection to their data and resources. Historically that investment has been very low - barely enough to ensure the continuity of operations. There are two perspectives that are generally taken into account when determining this. The first is the IT manager's perspective and determining what factors of the IT manager motivate them to invest in cybersecurity. These factors could be internal (attitude, self-efficacy, experience, and habits) or external, including the severity of past attacks this manager has experienced. The second perspective to consider is the financial perspective. According to Fernandez De Arroyabe, et. al. (2023) the financial perspective emphasizes the value of the investment in cybersecurity, meaning that the decision to invest is determined by the profit of the investment. The investment in cybersecurity is a decision that should be made by the organization in its entirety. All levels from senior management to the lowest cyber technology should be involved in this process.

Economics are not the only incentive that organizations look at when determining the use of cybersecurity within the organization. Wessels, et. al. (2021) argue that normative incentives can impact the use of cybersecurity within an organization as well. When decisions are made to invest in cybersecurity on the shared values and norms that are present for IT quality, a duty to contribute to society's safety, and a feeling of belonging with what everyone else is doing that would indicate a normative incentive. Those incentives include Historic incentives, Network incentives, or with competing solutions. Historic incentives, or decisions that are made to invest in cybersecurity based on past decisions and experiences, are another way that companies incentivize the use of cybersecurity. Wessels, et. al. (2021) stated that decisions made in the past to protect the organization's technology can impact current decisions and may even cause the company to act proactively to avoid past problems in the future. Network incentives are when decisions are made based on the organization's specific environment. These decisions are impacted by what other companies are doing or by the current international views on specific cybersecurity issues and is another way that cybersecurity investment decisions are made. For instance, if one Electronic Health Record (EHR) corporation has a data breach, it may drive consumers to a different company. This is not the case when a single corporate entity has captured a market (also known as monopoly) or if that industry has effectively regulated competing interests out of that field (as in regulatory capture). The feasibility of implementing specific measures in cybersecurity is another decision-making model. The final incentive model is with competing solutions and problems. Wessels, et. al. (2021) said that this is when decisions are made to invest in cybersecurity based on an organization's awareness of competitive cybersecurity issues and solutions that are complex in scope, benefit, and cost.

Fernandez de Arroyabe, et. al. (2023) stated that the investment in cybersecurity, regardless of incentive, is a strategic decision within the company. The decisions that are made must be based on a combination of cyber-capabilities of that entity, external threats that are likely to be faced, and attacks with the knowledge that attacks are becoming bigger with more sophisticated means of delivery, making them harder to mitigate. Having developed procedures, rules and resources for cybersecurity has a positive impact on investments in this area. Even though investment in cybersecurity is conditioned by the capabilities, competencies and attacks of a specific organization, the possession of developed cybersecurity procedures has reduced the severity of threats and positively impacted the entire company.

4. Network Considerations of Security Typology

Kalkman & Wieskamp (2019) state that the type of network that is being used is also impacted by the lack of a clear typology for cybersecurity in terms of challenges and outcomes of cyber intelligence. There are four network types (centralized, business, operational, and local) that all are impacted by cyber threats in different ways. However, with the frequency of change in the cyber threat landscape, it is nearly impossible for organizations to manage all of the threats on their own. Kalkman & Wieskamp (2019) argue that it is of utmost urgency to build a shared intelligence network that is collaborative. Collaboration is tricky because all of the different network types have unique characteristics which makes it both important to distinguish between network types when collaborating, and for those competing network types to collaborate. Top-down networks (centralized and business) generally have a harder time staying committed to other parties and seeing results in their collaborative efforts while bottom-up networks (operational and local) are typically less hierarchical and formal and yield more successful results. Regardless of network type, it is critical to improve cyber threat intelligence and create a unified effort against cybersecurity threats to protect global assets.

In regard to the typology of the network and cybersecurity both the demilitarized zone (DMZ) and firewalls were discussed during the scoping review. Hawrylak, et. al. (2019) stated that while the DMZ does provide high bandwidth and throughput connectivity between computing resources it generally has minimal security. Maintaining the confidentiality and integrity of data on the DMZ is challenging. Security solutions in this instance must balance the fast speed for data processing with strong integrity and confidentiality to prevent malicious use of computing resources. Whitelists and strong SSH (Secure Shell) keys can be shared between users on the same local computer in order to help manage security on the DMZ. Data encryption on the DMZ will allow data to be protected at rest as well. Hawrylak, et. al. (2019) argued that passive security tools are another way that DMZs can remain secure. When the DMZ is part of a cybersecurity strategy (in terms of network segmentation) it cannot and should not be viewed exclusively as a performance solution. According to Anwar, et. al. (2021) firewalls are the first line of defense for security in many cases. Firewalls are critical for securing sensitive data and act as a filter between outgoing traffic from the internal network and incoming traffic coming from outside the network. The effectiveness of the firewall is dependent on environmental conditions (including the layer that the firewall is at and the nature of the data being protected). Anwar, et. al. (2021) stated that the effective use of firewalls reduces the impact of cyber threats on a company. Even with the importance of these two specific pieces of technology, there is not an agreement about how they should best be secured in order to ensure the security of the network.

5. Impact of Behavior on Security Typology

One important component of any cyber-enabled network is that people will always be involved with the security of data. People are the most common threat to keeping data secure through phishing emails, supply chain threats, threat actors, and other emerging risks. People engage in risky behavior and that needs to be addressed when looking at bio cybersecurity.

According to Buckley, et. al. (2023) there are many factors that determine how people respond to phishing emails including sociodemographics, cyber security training, phishing email typology, information processing factors and faith in the institution itself. Buckley, et. al. (2023) state that there has been a rapid increase in the number of data breaches from phishing emails as a result of the increased amount of remote work being done from the COVID-19 pandemic. When secure email behaviors are being followed, there is early detection and abatement of cyber threats which reduces financial costs and disruptions to the workplace.

Supply chain risks are also impacted by human interaction. Simon & Omar (2020) stated that a lack of coordination along the supply chain can result in either underinvestment or overinvestment from specific nodes in that chain, but there is not a lot of coordination between nodes. Breaches in the supply chain can occur quickly but may take months or even years to detect and contain. This can cause disruptions in logistics, production, and operations and may result in a loss of data. This data loss can lead to additional costs to return the supply chain to its original state. Simon & Omar (2020) argue that it should be noted that cybersecurity for a supply chain cannot be an isolated view, there has to be awareness along the entire chain as to how to mitigate cyber risks.

Additionally, Ursillo & Arnold (2019) argue that it is of utmost importance to have proper IT procedures in place to mitigate the risks associated with human error. Ongoing education to employees should be a part of all organizations' risk management framework. Policies to help mitigate risk from employees could include mandatory use of passwords on all systems (with passwords being changed regularly), encrypted USB sticks, physical security of all equipment, and multifactor authentication if possible. Maalem, et. al. (2020) stated that human error is the cause of many new cyber-attacks. Threat actors are more sophisticated and can use advanced techniques to hack into systems through human-enabled actions. Kraemer, et. al. (2009) stated that problems with passwords in particular may impact security and should be addressed to improve cybersecurity measures. Human errors may also be related to organizational factors like communication, security culture and company policy - and these mistakes may be costly for organizations to recover from.

Maalem, et. al. (2020) argued that behavioral aspects of cybersecurity should be scrutinized to find a way to lessen the number of successful attacks. It is the job of the cybersecurity team to analyze security at all levels of the network and ensure defenses are present throughout. The threat to data (confidentiality) occurs when hackers attack databases, backups, application servers and try to gain system administrator access. The threat to integrity, or altering data, includes hijacking, changing financial data, stealing large sums of money and rerouting direct deposits, or damaging an organization's reputation. Denial of access to a system, or a threat to a system's availability, involves DDoS attacks and physical destruction of data. All of these have implications to

the biological field - including changing the deliveries of medical components, release of patient records, ransomware attacks on healthcare systems, or DDOS threats to publishing institutions (Mueller, 2021; Samori et al, 2022; Finch et al, 2023, Stephen, 2023). Weapons of influence (reciprocity, scarcity, authority, consistency, liking and consensus) are largely used by hackers and ignored by defenders which is a problem. If a standard topology existed for cybersecurity, this behavioral component could be included which would allow for all companies to explore human factors and how they are integrated with cybersecurity. The lack of cyber expertise within organizations that is argued by Kraemer, et. al. (2009) decreases the effectiveness of the cybersecurity policies that are in place. These cyber system vulnerabilities cannot necessarily be addressed by technical remedies, since the behaviors of individuals are what cause the problems to take place.

6. Limitations

A common admonition is not to miss the forest for the trees which compose it. Yet in an increasingly complicated threat environment, it is sometimes difficult to engage wholly with the staggering number of issues that can be present in a biology-linked cybersecurity environment. The hope of this work is to re-engage cybersecurity professionals in biologically related fields with the sum total of the meaning of and drives behind security. That being said, there is a single caveat. Sometimes institutions are named as too big to fail. Leonhardt reports that Equifax, a U.S. credit reporting agency, may be required to pay up to 700 million USD (2019). The EBITDA for Equifax in 2020 was 1.628 billion USD and in 2021 was 1.076 billion USD. The required costs of the changes required in IT infrastructure according to an Equifax press release are even less than the maximum legal penalty (2022). Thus, the lesson from legal authorities to corporate entities was that a large enough institution would not face consequences for their failures, and the message conveyed to users of the corporate entity was that their perception of cybersecurity competence was unimportant. The limits to this study and to government policy may be non-obvious - but in this case shows that there are certain institutions that may be not only resistant to change but can forestall what should be common sense regulations permanently. This grows ever more concerning when considering the impacts this phenomenon might have on a biologically relevant institution, such as a hospital network or government healthcare program.

7. Conclusion

According to Ursillo & Arnold (2019) protection from cybercrime should be a focus for all organizations and governments. The monetary and reputation risks are extremely high without proper cybersecurity procedures in place. Cybersecurity governance and management must be established for every organization, and as previously noted, there is not a lot of agreement about what a true cybersecurity typology should look like. Our literature review has found that throughout the three governance models and six broad categories of cybersecurity that the networks of threats that contribute to the perception of security is a worthy source of study that is in an interrelated relationship to the behaviors of the participants of the organization. The prime limitation of the typology of cybersecurity is that the economic costs of implementing best practices are not in line with the profit motive of a corporate entity, and some corporate entities are so large as to be beyond the ability of legal practices to effectively change. Thus, an issue that takes primacy is the reluctance of government organizations to make cyber policy a matter of jurisprudence. This is a consequential matter, as the current lack of government command of cybersecurity infrastructure often leaves government agencies “picking up the pieces” after an incident for private interests. Since many of these private interests currently are integral components of goods and services required for the daily operation of a nation (healthcare systems, research institutes, agricultural operations, etc.) the clock is ticking before this method of reactance to issues and misunderstanding of typologies leads to an incident which leaves needed components of daily life untended to.

References

Anwar, R. W., Abdullah, T., & Pastore, F. (2021). Firewall Best Practices for Securing Smart Healthcare Environment: A Review. Applied Sciences, 11(19), 9183. MDPI AG. Retrieved from http://dx.doi.org/10.3390/app11199183

Bevilacqua, S., Neira-Villena, J. E., & Valverde, M. (2022). La tecnología al servicio de la vigilancia y de la defensa de la vida. Estudios en Seguridad y Defensa, 17(33), 179-200.

Buckley, J., Lottridge, D., Murphy, J.G., and Corballis, P.M. (2023). Indicators of employee phishing email behaviours: Intuition, elaboration, attention, and email typology. International Journal of Human-Computer Studies. https://www.sciencedirect.com /science/article/abs/pii/S1071581923000022

Chappell, B. (2020, June 1). Protesting Racism Versus Risking COVID-19: 'I Wouldn't Weigh These Crises Separately'. Retrieved from https://www.npr.org/sections/coronavirus-live-updates/2020/06/01/867200259/protests -over-racism-versus-risk-of-covid-i-wouldn-t-weigh-these-crises-separate

Cremer, F., Sheehan, B., Fortmann, M., Kia, A., Mullins, M., Murphy, F. & Matarne, S. (2022). Cyber risk and cybersecurity: a systematic review of data availability. The Geneva Papers on Risk and Insurance - Issues and Practice, 47, 698–736. https://doi.org/10.1057/s41288-022-00266-6

Curwell, P. (Aug. 2022). Typologies demystified - what are they and why are they important? @Forwarned Blog. Retrieved from https://forewarnedblog.com/2022/08/20/typologies-demystified-what-are-they-and-why-are-they-important/

Duncan, S. E., Reinhard, R., Williams, R. C., Ramsey, F., Thomason, W., Lee, K., ... & Murch, R. (2019). Cyberbiosecurity: A new perspective on protecting US food and agricultural system. Frontiers in bioengineering and biotechnology, 7, 63.

Eggenschwiler, J. (2018). A Typology of cybersecurity governance models. St Antony's International Review, 13(2), 64–78. https://www.jstor.org/stable/26501049

Equifax, (9 Feb 2022). Equifax Delivers Record Revenue and Eighth Consecutive Quarter of Double-Digit Growth. Retrieved from: https://investor.equifax.com/news-events/press-releases/detail/1214/equifax-delivers-record-revenue-and-eighth-consecutive

Evangelista, S. (2020, April 30). COVID-19: Using wastewater to track the pandemic. Retrieved from https://medicalxpress.com/news/2020-04-covid-wastewater-track-pandemic.html

FDA Workflow Improvement Toolkit. (2019, December 12). Retrieved from https://www.hhs.gov/cto/projects/fda-workflow-improvement-toolkit/index.html

Fernandez De Arroyabe, I., Arranz, C.F.A., Arroyabe, M. F., and Fernandez de Arroyabe, J.C. (2023). Cybersecurity capabilities and cyber-attacks as drivers of investment in cybersecurity systems: A UK survey for 2018 and 2019. Computers & Security, 124. https://www.sciencedirect.com/science/article/pii/S0167404822003467

Friedersdorf, C. (2020, June 4). The Protesters Deserve the Truth About the Coronavirus. Retrieved from https://www.theatlantic.com/ideas/archive/2020/06/protests-carry-risk-even-when-theyre-justified/612652/

Finch, H., Affia, A. A., Jung, W., Potter, L., & Palmer, X. L. (2023, February). Commentary on Healthcare and Disruptive Innovation. In International Conference on Cyber Warfare and Security (Vol. 18, No. 1, pp. 77-84).Fi

Griffin, B., Alexander, K., Potter, L., & Palmer, X. L. (2023, March). Social-Engineering, Bio-economies, and Nation-State Ontological Security: A Commentary. In ICCWS 2023 18th International Conference on Cyber Warfare and Security (Vol. 42, No. 1, pp. 127-142). Academic Conferences and publishing limited.

Hawrylak, P.J., Louthan, G., Hale, J., and Papa, M. (2019). Practical cyber-security solutions for the science DMZ. Association for Computing Machinery, New York, 50,(1-6). https://doi.org/10.1145/3332186.3332213

Kalkman, J.P. & Wieskamp, L. (2019). Cyber intelligence networks: A Typology. The International Journal of Intelligence, Security, and Public Affairs, 21(1), 4-24. https://www.tandfonline.com/doi/full/10.1080/23800992.2019.1598092

Kacprzak, M., Attard, E., Lyng, K. A., Raclavska, H., Singh, B., Tesfamariam, E., & Vandenbulcke, F. (2022) Biodegradable Waste Management in the Circular Economy. DOI: 10.1002/9781119679523

Kraemer, S., Carayon, P., & Clem, J. (2009). Human and organizational factors in computer and information security: Pathways to vulnerabilities. Computers and Security, 28, 509-520.

Leonhardt, M., (2019), Equifax to pay $700 million for massive data breach. Here's what you need to know about getting a cut. CNBC.com. https://www.cnbc.com/2019/07/22/what-you-need-to-know-equifax-data-breach-700-million-settlement.html

Maalem Lahcen, R.A., Caulkins, B., Mohapatra, R. et al. (2020). Review and insight on the behavioral aspects of cybersecurity. Cybersecurity, 3. https://doi.org/10.1186/s42400-020-00050-w

McQuate, S. (2019, October 29). Popular third-party genetic genealogy site is vulnerable to compromised data, impersonations. Retrieved from https://techxplore.com/news/2019-10-popular-third-party-genetic-genealogy-site.html

Mueller, S. (2021). Facing the 2020 pandemic: What does cyberbiosecurity want us to know to safeguard the future?. Biosafety and Health, 3(1), 11-21.

Palmer, X. L., Powell, E., & Potter, L. (2021, June). Biocyberwarfare and Crime: A Juncture of Rethought. In European Conference on Cyber Warfare and Security (pp. 517-XIV). Academic Conferences International Limited.

Peccoud, J., Gallegos, J. E., Murch, R., Buchholz, W. G., & Raman, S. (2018). Cyberbiosecurity: from naive trust to risk awareness. Trends in biotechnology, 36(1), 4-7.

Potter, L. and Palmer, X.L., 2021, April. Human Factors in Biocybersecurity Wargames. In *Future of Information and Communication Conference* (pp. 666-673). Springer, Cham.

Reed, J. C., & Dunaway, N. (2019). Cyberbiosecurity Implications for the Laboratory of the Future. Frontiers in bioengineering and biotechnology, 7, 182.

Samori, I. A., Palmer, X. L., Potter, L., & Karahan, S. (2022, September). Commentary on Biological Assets Cataloging and AI in the Global South. In Intelligent Systems and Applications: Proceedings of the 2022 Intelligent Systems Conference (IntelliSys) Volume 3 (pp. 734-744). Cham: Springer International Publishing.

Simon, J. and Omar, A. (2020). Cybersecurity investments in the supply chain: Coordination and a strategic attacker. European Journal of Operational Research, 282(1), 161-171. https://www.sciencedirect.com/science/article/abs/pii/S037722171930757X

Stephen, S., Alexander, K., Potter, L., & Palmer, X. L. (2023, February). Implications of Cyberbiosecurity in Advanced Agriculture. In International Conference on Cyber Warfare and Security (Vol. 18, No. 1, pp. 387-393).

Stone, W. (2020, June 5). Tear Gassing Protesters During An Infectious Outbreak Called 'A Recipe For Disaster'. Retrieved from https://www.npr.org/sections/health-shots/2020/06/05/870144402/tear-gassing-protesters-during-an-infectious-outbreak-called-a-recipe-for-disast

Towie, M., & Gale, J. (2020, May 22). A Deadly Virus From Africa Is Killing Horses in Thailand. Retrieved from https://www.bloomberg.com/news/articles/2020-05-22/a-deadly-virus-from-africa-is-killing-horses-in-thailand

Ursillo, S., and Arnold, C. (2019). Cybersecurity is critical for all organizations - Large and small. International Federation of Accountants. https://www.ifac.org/knowledge-gateway/preparing-future-ready-professionals/discussion/cybersecurity-critical-all- organizations-large-and-small

Wessels, M., van den Brink, P., Verburgh, T., Cadet, B., & van Ruijven, T. (2021). Understanding incentives for cybersecurity investments: Development and application of a typology. Digital Business, 1(2), 1-7. https://www.sciencedirect.com/science/article/pii/S2666954421000132

Yi, Y., Lagniton, P. N., Ye, S., Li, E., & Xu, R.-H. (2020). COVID-19: what has been learned and to be learned about the novel coronavirus disease. International Journal of Biological Sciences, 16(

Assessment of Cyber Security risks: A Smart Terminal Process

Jouni Pöyhönen[1], Jussi Simola[1], Irfan Khan[2], Martti Lehto[1] and Syed Wali[2]
[1]University of Jyväskylä, Jyväskylä, Finland
[2]Texas A&M University, Galveston, TX USA

jouni.a.poyhonen@jyu.fi
Jussi.hm.simola@jyu.fi
irfankhan@tamu.edu
martti.j.lehto@jyu.fi
syedwali@tamu.edu

Abstract: In Finland, the connections to global maritime transportation logistics systems are an essential part of the national critical infrastructure. As a part of maritime logistics systems, the port's operations are important elements for global maritime traffic and the transportation supply chain. Digitalization of seaport services makes it possible to increase the efficiency of terminal systems in the logistic processes. At the same time, port logistic processes can notably reduce its CO2 emissions by optimizing port operations. The improvement of port processes relies very much on the development of Information and Communication Technology (ICT) and Industrial Control Systems (ICS) or Operation Technologies (OT) systems. In port environment there are parts that are controlled (ICS/OT) from the cyber environment but directly interact with the physical surroundings. These are called Cyber-Physical Systems (CPS). In this environment, the cyber security aspects of the port logistic need to be addressed. In Finland, the Port SMARTER research program has been on the way since 2021. The aim of the program is to create port services within new technology solutions, and that way improve cargo and people flows while improving the experience for all stakeholders. However, this development increase also complicated system dimensions in the use of ports and makes port operations complex systems of systems environment characterized by a conglomeration of interconnected networks and dependencies. This paper describes a practical approach to risk assessment work regarding the SMARTER research case. It provides a comprehensive cyber security investigation approach to port operations at the system level. In risk assessment work, the paper emphasizes the importation of description of probabilities to defend the system element against estimated probabilities of cyber-attacks at all parts of port processes. The findings of the study are related to the comprehensive cyber security architecture of the SMARTER research goals. The following research interests are related to the issue: "How a comprehensive cyber security investigation can be conducted in smart ports operations?" This paper emphasizes cyber security risks assessment work should be covered from services for operation, information flows in and between systems, as well as electricity supplies to achieve holistic risks assessment in the smart terminal process.

Keywords: Maritime Logistic, Smart Terminal Process, Cyber Security Management, Risk Assessments

1. Introduction

ENISA port cybersecurity report (2019) emphasizes the importance of maritime transport systems for the European Union economy. It means the activity that relies on more than 1 200 seaports within the European Union. Each of them is with a different organization, interests, challenges, and activities. (ENISA, 2019)

International and national maritime transportation systems are essential parts of critical global infrastructures. Digitalization and increased levels of autonomy in logistic transport chains are expected to take leaps forward in the coming years. In Finland Smart Terminals (SMARTER) research project consists of ports digitization by the end of 2023. The mission of SMARTER is to create replicable models for digitalization, service innovation, and data usage and sharing in the harbor environment and prepare for the future by taking steps towards smart and autonomous maritime transportation. The project goals are conducted to the reduction of emissions by optimizing harbor operations and improving cargo, and people flow while improving the experience for all stakeholders. (DIMECC, 2020)

The structure of the project has been planned to have three use cases. The use cases are ship turnaround, truck traffic, and passenger flow. Use cases are designed to support one another, and there are linkages between the use cases. The applied research work is organized into five work packages, including cyber security research actions in Work Package 4. (DIMECC, 2020)

Modern society depends entirely on a cyber environment that provides dynamic services. The port cases digitalization means the development of Information and Communication Technology (ICT), Information Technology (IT), and Industrial Control Systems (ICS) or Operation Technologies (OT) solutions. The maritime ports are digitalized System of Systems (SoS) where system-level threats are needed to be coordinated like hybrid responses. Therefore, a system of systems-level research view is also needed. Thus, it is necessary to

address the relevant cyber safety aspects of the overall maritime solutions. In any cyber environment, it is crucial that there are trustable information networks. In addition, the usability, reliability, and integrity of systems data need to be high within the operating environment, where cyber security risks are continuously being highlighted by the threatening scenarios posed by the digital world. ENISA Threat Landscape Report 2016 emphasizes all elements covered within an attack on a business process. It means that not all artifacts/components used are IT-related; there are steps/procedures used within an attack, that are performed by just having knowledge or information about the details of the business process at stake (ENISA, 2017).

In Finland, the Sea4Value research program has been established to increase digitalization for fairway and smart port (SMARTER) services. This paper follows our four previous research papers in smart port study concerning cyber risk challenges in maritime transportation, basic elements of cyber security, cyber threat analysis, and cyber threat impact evaluation. This paper counts on the importation of system of systems description of the port process and specifies the comprehensive cyber security architecture by exploiting the risks assessment process for cyber security measures for the SMARTER project. The whole Sea4Value cyber security study is a set of papers of our research team and together is also an example of soft systems research methodology used to improve cyber security management in a system of systems environment.

2. Risk assessment for Smart Terminal

The case of port in maritime transportation system includes processes to produce all needed services, like ship approached from an open sea via fairway to berthing to a pier, general port services, port logistics, and connections to land transportation. There are identified the key elements associated with processes used in the smart terminal in our previous paper, "Basic Elements of Cyber Security for a Smart Terminal Process" (2023). These are Activities, Stakeholders, Organizational relationships, Security dimensions, Security capabilities, and Criteria (Pöyhönen, Simola, Lehto, 2023). It is also evident that this entity needs communication systems within process elements and electricity systems to support the functions of processes. In all cases of port processes, the information requirements and the amount of information needed are related to the reliability of safety and security services. Cyber security awareness and information should cover all process elements. Figure 1 presents these processes (Simola & Pöyhönen, 2022).

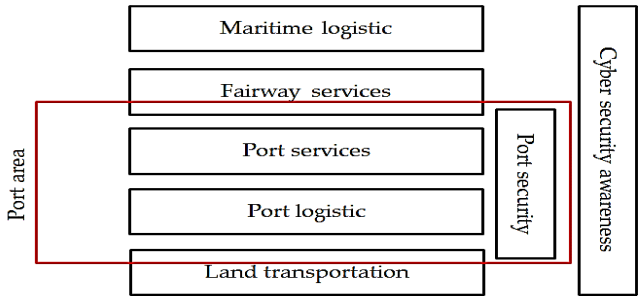


Figure 1 Port Processes

A block diagram is a useful way to describe cyber security's main dimensions for a SoS approach in terms of general smart port elements. Figure 2 illustrates the functional blocks of port operations (Simola & Pöyhönen, 2022). Communication lines between blocks are needed to identify and understand the basic architecture of its cyber security management.

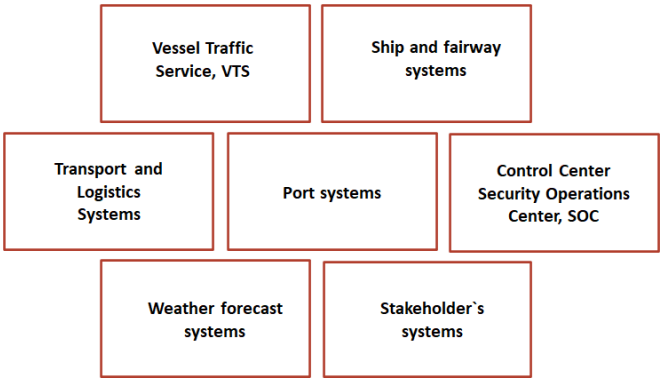


Figure 2. Block diagram for the Port Operations

The functional blocks of the port make it possible to produce in process context necessary services to port operations. In addition to information flows that are needed to produce services, the electricity supply for the port processes is also important to secure. According to our previous articles (Pöyhönen & Lehto, 2017; Pöyhönen, 2022), the concept of national critical infrastructure can be simplified into three essential layers. At the base layer is the electric network, above that is the data/information transmission layer and above theirs are services. In SoS thinking, all layers together is a comprehensive cyber entity. That is the case in port operations also. The function of an entity depends on the process reliability and continuity of these layers. It can be considered that port services, according to block diagram systems and communication between them, are the main systems. All structures of electric power systems are the support systems for the port services. Thus, these systems are security dimensions to be considered as part of cyber security estimation.

Cyber security estimation is very much a question of correctly implemented and appropriately functioning risk management of an organization or process. Assessment of cyber risks is the most important process behind all digital security, including data security, privacy protection, and business continuity management. Risk management is increasingly important as the need for improving the various areas of security has increased. The need for improvement has arisen from the digitalization of operations, the possibilities offered by new technologies, and the new threat and risk types that have evolved rapidly. Without an appropriately functioning risk management, the organization may not be able to recognize the significant threats that could prevent the achievement of its objectives or that are related to its daily operations and will not be able to control these threats (Ministry of Finance, 2017).

Risk management is also an excellent tool for the organization when it develops the processes, actions, and services to improve its security. Risk management helps achieve cost-efficiency, allowing development measures to be targeted at matters that have a significant impact on decreasing the probability or mitigating the impact of a recognized threat. In addition to risk management, the guideline discusses the recognition of opportunities. The failure to take advantage of opportunities could pose a threat to improving the organization's operations or achieving its targets, for example. Increased digitalization of operations is a good example of such an opportunity; it should be seen as an important player in developing operations. However, the threats related to digitalization must also be recognized.

Following figure 3 illustrates the risk assessment work as a part of the holistic cyber security architecture process.

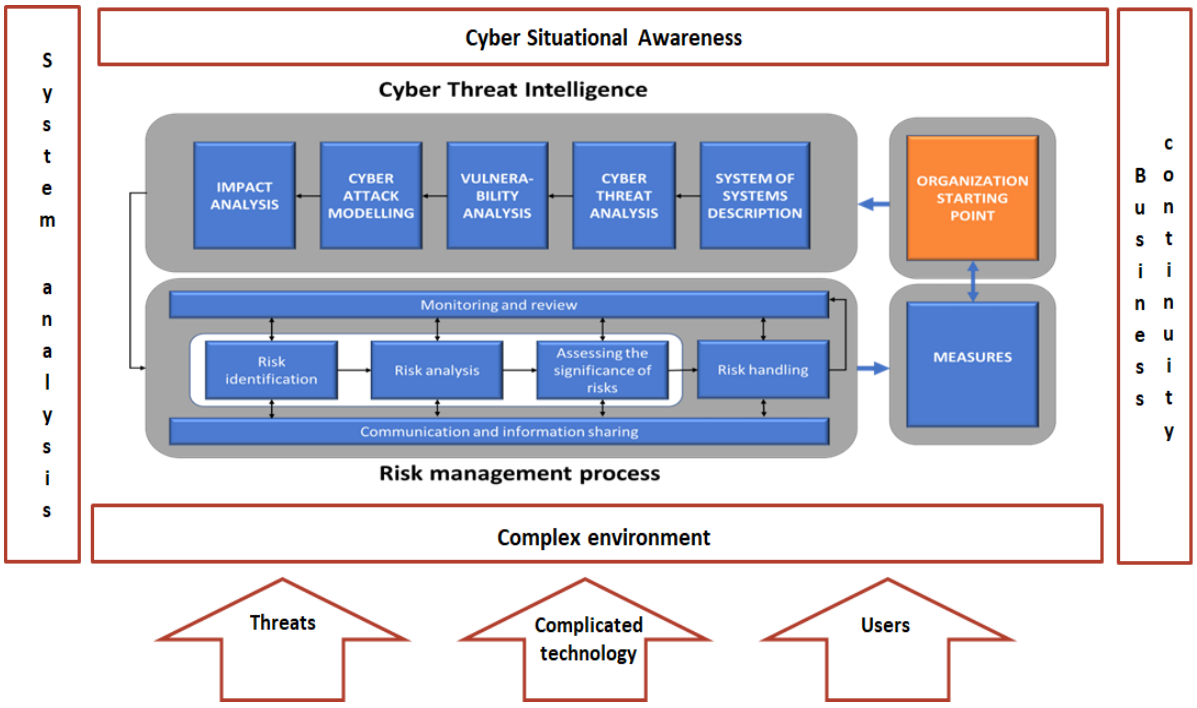


Figure 3. Risk assessment as a part of the holistic cyber security architecture process

3. Research environment for the risks assessment process

This section is based on authors work used research framework in Sea4Value research program and it also underlines our definitions of relevant smart port elements for cyber security investigation. Firstly we have a description of a research framework for cybersecurity study of the system of systems in our previous papers concerning maritime cyber security outcomes. It is a combination of a five-layer cyber structure of an organization and the maritime security management system with aspects of cybersecurity and the block diagram of a process of main maritime systems. In addition to this, an organization's cybersecurity management requires comprehensive awareness at the system level. The awareness of an organization and decision-makers can be seen as a system-level awareness arrangement. It is possible to integrate an organization's three decision-making levels into a five-layer cyber structure (Figure 4) in order to have a comprehensive system view of that organization's cybersecurity environment (Pöyhönen & Lehto, 2020).

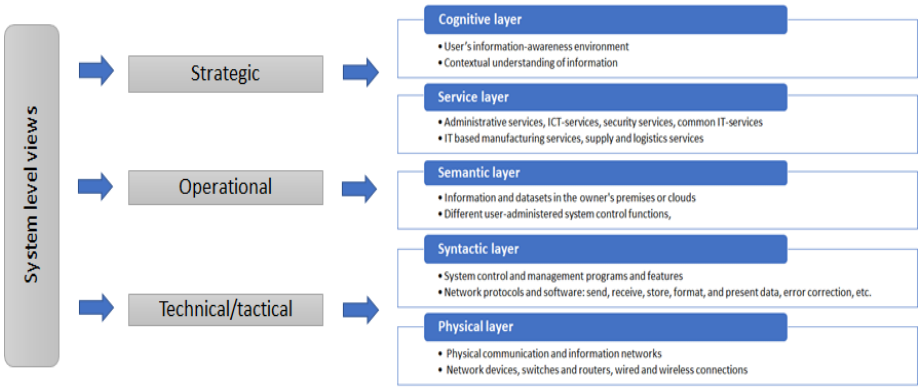


Figure 4. System-level view of organizational cybersecurity (Pöyhönen & Lehto, 2020)

Secoudly of our previous work, the paper "Basic Elements of Cyber Security for a Smart Terminal Process" (2023) presents the findings of cyber security research on the SMARTER project (Pöyhönen, Simola & Lehto, 2023). The key elements associated with operations and processes used in the SMARTER research areas are listed in the headlines level as follows: Activities, Stakeholders, Organizational relationships, Security dimensions, Security capabilities, and Criteria. Figure 5 (Pöyhönen, Simola & Lehto, 2023) illustrates the cyber security elements and features of this study . To cover all these, the result can be called a Smart Port, Cyber Security Management System (PortCSMS). In this case, these key elements and features can be considered root definitions of relevant systems for cyber security investigation (Checkland, 1981).

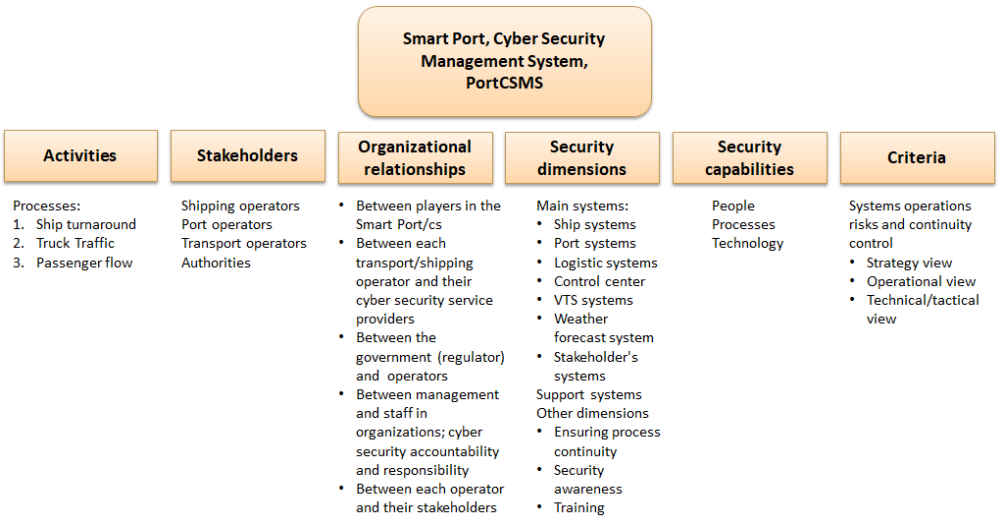


Figure 5. Elements of smart port cyber security management, Port CSMS.

In the SMARTER research project is acknowledged that there is a need for new technologies and solutions that are necessary to tackle the challenges set by the use cases (DIMECC, 2020). It is identified as a set of technologies and solutions that belongs to the terminology of Industry 4.0, like big data, data lake, data analytics, information fusion, AI, 5G, IoT, edge processing etc. It is reconditioned in the article of "Cybersecurity Risks and Automated

Maritime Container Terminals in the Age of 4IR (Fourth Industrial Revolution)" by Peter Beaumont (2018) that "the risks derived from the use of technology associated with the Fourth Industrial Revolution (4IR) are both real and dangerous unless appropriate control measures are implemented" (Beaumont, 2018). It is also important to recognize and knowledge legacy systems in use with a different timeframe of technology solutions. In that sense, the SoS environment of the port process is complicated from a technological point of view. In this SMARTER cyber security research, either new and legacy technologies should be covered in bout security dimensions, main and support systems of the case.

The main systems of this study are described in a block diagram (Figure 2). It describes the cyber security main dimensions of the SoS approach in port operations (Simola & Pöyhönen, 2022). The figure shows the functional sides of the port that are needed in order to understand the basic architecture of its cyber security management. Communication and relationships between the blocks in the port systems are one of the key features of its operational processes. Next, they are emphatical to be studied in this research process (Pöyhönen, Simola & Lehto, 2023).

The support systems are related to national critical infrastructure; electric power systems, information networks and other services that are needed to support systems for the port services and thus also taken as security dimensions as a part of cyber security estimation.

The concept of national critical infrastructure can be simplified in accordance with three layers of services. At the base layer is the electric network, above that, is the data transmission layer, and above theirs are services. The base layer of smart terminals is crucial from the perspective of security and operational integrity. Major port operations can be easily disrupted by a power outage in smart ports. An adversary can easily achieve a power outage by manipulating artifacts or parameters of the control system. This outage can be regarded as a Denial of Service (DoS) attack because its primary outcome is to shut down the power of essential port services. Unavailability of power for a long period can produce serious consequences. For instance, the main concern of a port specialized in cargo will be related to cargo maintenance and storage. Forexample, refrigerated containers consume up to 40% power of a terminal for storing consumer goods (FCH, 2017; Duin, Geerlings, Tavasszy, & Bank, 2019). However, depriving them of their essential power requirement for a longer period may result in content spoilage and monetary loss to stakeholders.

The second critical layer of smart ports is the data transmission layer that allows human-to-machine, machine-to-machine, and human-to-machine connections. It transfers data between different components of a port and is extremely critical for a series of services. Port ICT system usually hosts databases, storing vessel and port information, and Terminal Operating System (TOS), performing critical stowage planning operations. All port ICT systems having network interfaces for external connections must be thoroughly assessed for potential vulnerabilities and their impact. Since an adversary can easily disrupt port operations and availability of TOS by exploiting vulnerabilities of the data transmission layer, rigorous risk assessment and risk management must be performed to protect the ports' operational integrity.

On top of the port infrastructure, there is a service layer that utilizes physical assets and communication channels for performing essential maritime operations (DNV GL, 2015). Smart ports are enormous implementations of software, hardware, and communications channels (Gary C. Kessler & Steven D. Shepard, 2020). The potential for attacks on port hardware or software is always present and evolving with respect to time. Emerging cyber threats with an increased dependency on vulnerable communication channels have jeopardized all stakeholders, making as like Intrusion Detection Systems (IDS) the essential network security requirement. However, the introduction of adversarial attacks in the cyber domain highlights the need to upgrade these existing systems because they can be exploited by modern attack vectors (Wali & Khan, 2021). Now an adversary can hack or manipulate sensor data to disguise a cyberattack as legitimate network traffic. In addition, they could sorely disrupt vessel traffic by sending false tidal window messages, berthing data, or clearance time to enter the port. Any of these scenarios could disrupt port operations, potentially for days at a time (Kessler & Shepard, 2020).

4. Risks assessment principles and estimation process

The information flows use legacy as well as new technologies, and electric supply chains are very important parts of cybersecurity risk analysis work from the port service point of view. The recognition of these technologies enables us to identify different functions at the system level, carry out risk assessments and identify their residual risks with sufficient accuracy. In the same way, the dependencies of different systems need to be considered, and based on these dependencies, security and cybersecurity risks need to be identified. In our

previous papers, we have presented a probability approach to cyberattacks versus a probability to defend attacks and, in the end, to evaluate cybersecurity risks related to the relevant service operations.

A cyber-threat model captures information about potential cyber threats against a system, an enterprise, an SoS, a region, or a critical infrastructure sector. A cyber-threat model can serve as a basis for a variety of tasks in different scopes. Comprehensive cybersecurity needs a wide scale of analysis of a system of systems (or sub-system) against a set of threat events. It can often be impractical, and, in that sense, analysis of SoS could rely on the development and use of threat scenarios. A threat scenario could include the picture of a potential threat and the result of harmful consequences (Bodeu & McCollum, 2018). This has been part of our work in previous papers in this study.

The Bayesian attack graph model has been studied in several areas of security risk management. The paper "Dynamic Security Risk Management Using Bayesian Attack Graphs" (Poolsappasit, Dewri & Ray, 2012) proposes a risk management framework using Bayesian networks in order to quantify the chances of network compromise at various levels of system constructions. In the same sense, various threat risk analysis schemes have been developed to recognize the attack and implement the security safeguards to protect the system asset from cyber-attacks (Wang & Liu, 2014). Attack trees (AT) technique plays an important role to investigate the threat analysis problem to known cyber-attacks for risk assessment. An attack graph is based on a probabilistic metric model and can be used to quantify the cybersecurity issues of an SoS environment.

In our paper "Assessment of cybersecurity risks - Maritime automated piloting process" (2022), an attack tree graph is used to represent the relationship between threat and defense actions in the relevant case process. At the SoS level, it is more than a metric model, a way of thinking, because there are many layers in the system configuration. Exact quantitative probability calculation is therefore complicated, and results can be inaccurate. According to our experiences, we are familiar with the attack tree graph as a tool for risk assessment. The result of this represents the likelihood of an attack against the likelihood of defense against attacks. The final probability of success of defense measures versus attacks will be estimated, and the most serious attacks will be recognized and prioritized. This probability evaluation (not exact calculation) work is proposed to be done by cybersecurity experts by utilizing all relevant information available from the cybersecurity features of the used technology in the current case and information from stakeholders' capabilities to have defense measures. In this sense, we have been to propose the use of the Delphi method principle to make relevant threat analysis and risk level estimations from the systems. In the Delphi methode we need to use several estimation cycles in order to have sufficient results. The recommendation is three or more. It is a useful way of thinking about likelihood and probabilities at the system level of a process or an organization. (Pöyhönen & Lehto, 2022)

The risks probability estimation can be extended to the system level as described below. The National Institute of Standards and Technology (NIST) released recommendations as "Framework for Improving Critical Infrastructure Cybersecurity" (2018) for owners and operators of critical infrastructure to help them identify, assess, and manage cyber-risks. The Framework Core part of the guidance has a set of cybersecurity activities, outcomes, and informative references that are common across sectors and critical infrastructure. Elements of the Framework Core provide detailed guidance to help an organization align and prioritize its cybersecurity activities with its business/mission requirements, risk tolerances, and resources. The Framework Core consists of five concurrent and continuous Functions: Identify, Protect, Detect, Respond, Recover. There is mentioned that "The Framework is not a one-size-fits-all approach to managing cybersecurity risk for critical infrastructure" (NIST, 2018).

The probabilistic success of attacks, P(t), against the defense of system x can be evaluated and calculated as follows, adapting the principle in "Threat Analysis of Cyber-Attacks with Attack Tree+" (Wang & Liu, 2014, mod.)

$$P_{Ax}(t) = P_A P_{D`} = (P_{SOT})(1 - p_P(t))(1 - p_D(t))(1 - p_M(t))(1 - p_R(t)) \quad (1)$$

where, as a function of time t, a successful attack against a system x, P_{Ax} has attack identify for success probabilities P_A (P_{SOT}: Strategy level S; Operational level, O; Technical/tactical level, T) reduced by a defending mechanism, $P_{D`}$: Protection P, Detection D, Countermeasure M and Recovery R having the respective success probabilities p_P, p_D, p_M and p_R. (Pöyhönen & Lehto, 2022)

5. SMARTER cyber security risks assessment

Conclusively, all layers of smart terminals in the system of systems thinking can position their own strategic role and identify their operation as part of an entity whose other parts depend on reliably functioning of the three

layers of national critical infrasrtucture. This also facilitates the identification of cyber dependencies within the layers so that they can be secured with the most efficient and practical measures. For that reason, risk assessment work should be covered all those levels and related elements whit in a five-layer cyber structure as well as all organizations decisions levels. (Pöyhönen & Lehto, 2017; Pöyhönen, 2022). However, maritime cyber risk assessment is relatively new (IMO, 2017), and it has been highlighted as an emerging need of the maritime sector by several nations over the last few years (Wilshusen, 2015). Since all layers of smart terminals are interconnected and function together to achieve a specific task, conventional risk assessment models are inaccurate in mirroring the actual dynamics of port operations (Polemi, Ntouskas, Theoharidou, & Gritzalis, 2013; Tam & Jones, 2019). The static nature of conventional approaches cannot examine the cyber-physical aspect of port operations, as unique, physical accidents can take place due to cyber intrusions or vice versa. Risk assessment of each layer separately cannot address emerging threats associated with the interconnection of layers. Therefore, the risk assessment process of smart terminals must consider a multi-dimensional model of port operations, such that all layers and cyber-physical elements of port operations must be included. This multi-dimensional risk assessment process, shown in Figure 3, consists of the following four steps:

a. **Risk Identification:** All possible sources of cyberattacks, vulnerabilities, their causes, and potential impact are identified at this stage of risk assessment. Therefore, in the case of smart terminals, Grid vulnerabilities and possible cyber-physical attack vectors in the communication and port services layer must be identified at this phase. Moreover, existing safeguards (intrusion detection system, prevention, or cyber-defensive strategies) capable of reducing or mitigating the consequences of any hazardous event must be identified.
b. **Risk Analysis:** The frequency and consequences of any identified risk is determined at this phase. A hazardous event may have multiple consequences, including unavailability of port services, downtime of the power utility, and failure of the communication system. Consequences should not be limited to any individual layer because problems can traverse from one layer to another. For instance, intrusion in the communication layer may hijack port operations. Therefore, risk analysis should consider all potential consequences.
c. **Risk Evaluation.** Evaluation of identified risks and their potential consequences is performed at this phase. Depending on the severity and impact, a quantitative value is assigned to every possible risk. In the case of smart terminals, every risk ranging from the base layer (electrical system) to the port services will be assigned a value. This value depicts the severity of all possible risk factors.
d. **Risk Handling.** Risk handling utilizes the quantitative data produced from the risk evaluation phase. This data helps the stakeholders to develop and implement an appropriate mitigation plan and invest in cybersecurity. A wide range of risk handling strategies can be selected at this phase, including physical system upgradation, implementation of the automated intrusion detection system, and integration of AI-driven cyber defensive strategies.

The proposed risk handling plan is then evaluated against financial feasibility, and only those plans are selected by stakeholders that balance costs and efforts of implementation against obtained benefits. Moreover, periodic monitoring and review of the risk assessment process must be carried out to incorporate defensive measures against emerging cyber threats and vulnerabilities within the cyber-physical layers of smart terminals.

6. Conclusion

The Finnish smart port research program concerns port services for ship turnaround, truck traffic, and passenger processes. It is called SMARTER, and the research activities are related to the digitalization survey with the use of new technologies in order to achieve the main objectives, the reduction of emissions by optimizing port logistics, and is to enable exceptional flow and experience for the passengers and cargo.

Smart ports are enormous implementations of hardware, software, and communications, including the development of new apps. The potential for attacks directly on physical surroundings, CPS, via hardware or software is always present and evolving. All emerging vulnerabilities associated with the interconnection of ICT and ICS/OT layers in smart terminals must be thoroughly assessed. The vulnerable cyberspace and systems of future ports pose a serious threat to both the economy and security as cyberattacks continue to become more sophisticated and more difficult to detect in real-time. Therefore, dynamic risk assessments of port infrastructure, encompassing all layers and components of the port, must be carried out to improve cyber resiliency and operational integrity in the smart terminals.

This paper provides a research approach for needs of cyber security risks assessment work of the system element of the smart terminal process. The research approach includes the system of systems thinking. The risk findings of the study are related to the coverage of port services, information flows in and between systems as well as electricity supplies in the port process.

References

Beaumont, P., 2018. Cybersecurity Risks and Automated Maritime Container Terminals in the Age of 4IR [Fourth Industrial Revolution], Chapter in 'Handbook of Research on Information and Cyber Security in the Fourth Industrial Revolution', pp. 497–516, https://doi.org/10.4018/978-1-5225-4763-1.ch017

Bodeu, D. J. & McCollum, C. D., 2018. System-of-systems threat model. The Homeland Security Systems Engineering and Development Institute (HSSEDI) MITRE: Bedford, MA, USA

Checkland, P., 1981. Systems Thinking, Systems Practice, John Wiley & Sons, Ltd., 330 s.

de la Peña Zarzueloa, I., Soeanea M. J. F. & Bermúdez, B. L., 2020. Industry 4.0 in the port and maritime industry: A literature review. Journal of Industrial Information Integration 20 (2020) 100173.

DIMECC Oy, 2020. DIMECC Sea4Value/Smart Terminals (SMARTER). Project proposal for One Sea – autonomous maritime ecosystem.

DNV GL., 2015. Ship connectivity . Strategic Research & Innovation Position Paper

Duin, J., Geerlings, H., Tavasszy, L. & Bank, D., 2019. Factors causing peak energy consumption of reefers at container terminals. Journal of Shipping and Trade.

ENISA, 2017. Threat Landscape Report 2016, 15 Top Cyber-Threats and Trends.

ENISA, 2019. PORT CYBERSECURITY. Good practices for cybersecurity in the maritime sector. NOVEMBER 2019. https://www.enisa.europa.eu/publications/port-cybersecurity-good-practices-for-cybersecurity-in-the-maritime-sector

FCH, 2017. Development of Business Cases for Fuel Cells and Hydrogen Applications for Regions and Cities. Brussels.

International Maritime Organization, IMO, 2017. Guidelines on Maritime Cyber Risk Management. MSC-FAL.1/Circ.3 5 July 2017

Kessler, G. C. & Shepard, S. D., 2020. *Maritime Cybersecurity: A Guide for Leaders and Managers* .

Ministry of Finance, 2017. Risk management guideline. Kimmo Rousku (editor). Ministry of Finance publications 22/2017. Finland.

National Institute of Standards and Technology, NIST, 2018. Framework for Improving Critical Infrastructure Cybersecurity, April 16, 2018

Polemi, D., Ntouskas, T., Theoharidou, M. & Gritzalis, D., 2013. S-Port: Collaborative Security Management of Port Information Systems. Conference: Information, Intelligence, Systems and Applications (IISA).

Poolsappasit, N., Dewri, R. & Ray, I., 2012. Dynamic Security Risk Management Using Bayesian Attack Graphs. IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 9, NO. 1, JANUARY/FEBRUARY 2012

Pöyhönen, J. & Lehto, M., 2017. Cyber security creation as part of the management of an energy company. ECCWS 2017: Proceedings of the 16th European Conference on Cyber Warfare and Security (pp. 332-340). Published by Academic Conferences and Publishing International Limited. Reading. UK.

Pöyhönen, J. & Lehto, M., 2020. Cyber security: Trust based architecture in the management of an organization security. The proceedings of the 18th European Conference on Cyber Warfare and Security ECCWS2020, 25-26 June 2020, University of Chester, UK, pages 304-313

Pöyhönen, J. & Lehto, M., 2022. Assessment of cybersecurity risks - Maritime automated piloting process. The proceedings of the 17th International Conference on Cyber Warfare and Security. State University of New York at Albany Albany, New York USA 17-18 March 2022. pp 262-271.

Pöyhönen, J., 2022. Cyber Security of an Electric Power System in Critical Infrastructure. Cyber Security, Critical Infrastructure Protection. Martti Lehto Pekka Neittaanmäki Editors. Springer. Computational Methods in Applied Sciences. Volume 56. Chapter 9. (pp. 217-254). ISSN 1871-3033. ISBN 978-3-030-91292-5 ISBN 978-3-030-91293-2 (eBook) https://doi.org/10.1007/978-3-030-91293-2

Pöyhönen, J., Simola, J. & Lehto, M., 2023. Assessment of cyber security risks – Smart terminal process. The proceedings of the 22nd European Conference on Cyber Warfare and Security. ECCWS-2023

Simola, J. & Pöyhönen, J., 2022. Emerging cyber risk challenges in maritime transportation. Proceedings of the 17th International Conference on Information Warfare and Security, 2022. pp 306-314

Tam, K. & Jones, K., 2019. Situational Awareness: Examining Factors that Affect Cyber-Risks in the Maritime Sector. International Journal on Cyber Situational Awareness

Wali S. & Khan I., 2021. Explainable AI and random forest based reliable intrusion detection system.

Wang, P. & Liu, J. C., 2014. Threat analysis of cyber-attacks with attack tree+. Journal of Information Hiding and Multimedia Signal Processing, 5(4).

Wilshusen, G., 2015. Maritime critical infrastructure protection: DHS needs to enhance efforts to address port cybersecurity. GAO-16-116T.

Governance and Management Information System for Cybersecurity Centres and Competence hubs

Jyri Rajamäki and Janne Lahdenperä
Laurea University of Applied Sciences, Espoo, Finland

jyri.rajamaki@laurea.fi
janne.lahdenpera@student.laurea.fi

Abstract: Information sharing allows organizations to leverage the collective knowledge, experience, and analytical capabilities of their sharing partners in a community of interest. Sharing information is made easier with the help of a suitable information system. The DYNAMO project (10/2022-9/2025) creates tools for the cyber situational picture to support decision-making. One task of its mode of operation is to continue the development of the assets designed in the ECHO project (2/2019-2/2023). This article examines the design of ECHO's governance and management information system and how it can be applied to support the organisational processes and information-sharing needs of the collaborative network as a part of the DYNAMO project.

Keywords: Cybersecurity centre, Competence hub, Governance model; Data governance; Management information system

1. Introduction

The European Cybersecurity Competence Centre (ECCC) increases Europe's cybersecurity capacities and competitiveness. Each Member State should nominate one National Coordination Centre (NCC) and the Community formed from this would involve a large, open, and diverse group of actors involved in cybersecurity technology. ECCC will work together with NCCs to build a strong cybersecurity community. ECHO (2022), CONCORDIA (2022), SPARTA (2022), and CyberSec4Europe (2022) were the EU pilot projects launched to prepare the European Cybersecurity Competence Network that will be achieved through multiple different concepts, such as a governance and management model.

Cooperation and information sharing between different information security operators are based on heterogeneous data, models, and forecasts. The management of the information flows is an important area of successful governance of the European Cybersecurity Competence Network. The scope of DYNAMO (2023) is to generate a cyber situational picture for decision support. The DYNAMO approach is based on the refinement of the existing solutions of the ECHO project. This design science research (DSR) deals with the development of governance and management information systems (GMIS) for cybersecurity centres and competence centres. The extension of the ECHO GMIS system would enable information sharing and cooperation within the network. This study focuses on the conception, structure, and application of information-sharing and management systems for cybersecurity centres and competence hubs where the ECHO GMIS is based.

This design science research (DSR) follows the DSR framework and the DSR checklist questions (CQ) (Hevner & Chatterjee, 2010). These DSR CQs are mapped to this paper as follows:

- *DSR CQ1 - What is the research question (design requirements)?*
  The main research question: *What kind of information system best supports the operation of a collaborative network organization?*
- *DSR CQ2 - What is the artifact? How is the artifact represented?*
  Section 4 describes the platform of the information system and its most important functions.
- *DSR CQ3 - What design processes (search heuristics) will be used to build the artifact?*
  Section 4 presents the heuristic design cycles of the artifact.
- *DSR CQ4 - How are the artifact and the design processes grounded by the knowledge base? What, if any, theories support the artifact design and the design process?*
  Section 2 deals with the knowledge base of this study.
- *DSR CQ5 - What evaluations are performed during the internal design cycles? What design improvements are identified during each design cycle?*
  Section 4.4 discusses the self-evaluation of the artifact.
- *DSR CQ6 - How is the artifact introduced into the application environment and how is it field tested? What metrics are used to demonstrate artifact utility and improvement over previous artifacts?*
  Section 5 deals with the implementation of the artifact.

- *DSR CQ7 - What new knowledge is added to the knowledge base and in what form (e.g., peer-reviewed literature, meta-artifacts, new theory, new method)?*
  Section 'Discussions' discusses these items.
- *DSR CQ8 - Has the research question been satisfactorily addressed?*
  'Conclusions' answers to the research question.

2. Groundings to the Knowledge Base

2.1 Governance and Management Information Systems

Organization's governance function is responsible for determining its strategic direction. Chartered Governance Institute UK & Ireland (2022) defines corporate governance as "the system of rules, practices and processes by which a company is directed and controlled." They also add that corporate governance is a toolkit with which management and board are able to deal with the challenges of running a company more effectively. One function of corporate governance is ensuring that businesses have appropriate controls and decision-making processes in place so that the interests of all stakeholders are balanced. Corporate governance is also very important when meeting the requirements of laws and regulations such as the General Data Protection Regulation (GDPR) or the Network and Information Security (NIS) Directive.

Information governance involves the overall strategy for information, it handles information in the context of policies, systems, people, and processes. Utilizing its different elements can help us to create and maintain applicable policies and procedures that help meet the requirements for our data privacy. Information governance balances the risk that information presents with the value that information provides. Issues covered by information governance include information security and protection, compliance, data quality, data governance, electronic discovery, risk management, privacy, data storage and archiving, knowledge management, business operations, and management, audit, analytics, IT management, master data management, enterprise architecture, business intelligence, big data, data science, and finance.

The management functions of an organization take the strategic direction and translate it into actions that bring the organization closer to achieving its strategic goals. A management information system (MIS) is an information system used for decision-making, and the coordination, control, analysis, and visualization of information in an organization. The study field of MIS involves people, technology, and processes in an organizational context. In corporations, the goal of the use of MIS is to increase the value and profits of the business through IT tools used to support processes, operations, and intelligence. Information sharing is considered one of the most important ways to increase organizational efficiency and performance (Yang & Maxwell, 2011).

2.2 Information System Design Theories

The main function of any implemented information system is to improve the efficiency, effectiveness, and cyber resilience of the organization where it is implemented. For designing information systems with design science, Hevner, et al. (2004) present seven guidelines to follow. These guidelines are not to be adhered to too literally, but during the design process, each of them should be assessed at least somewhat. The guidelines are:

1. Design as an Artifact – The outcome of the research should be an IT artifact that focuses on a specific problem related to the organization.
2. Problem Relevance – It is important for information system research to address the problems that arise in the environment it is applied.
3. Design Evaluation – To demonstrate the quality of the design artifact, well-executed evaluation methods should be used.
4. Research Contributions – It is important to provide clear contributions in the area of the research artifact.
5. Research Rigor – The application of rigorous methods is required in both the construction and the evaluation of the designed artifact.
6. Design as a Search Process – The target of finding an effective solution to a problem is an inherently iterative search process.
7. Communication of Research – It is important to present the research results in a form that is easily understood both by management-oriented and technology-oriented audiences.

When designing an information system, it is also very beneficial to define different roles and responsibilities within the system right from the beginning. One way to accomplish this is to use a responsibility assignment matrix, also known as RACI matrix (RACI Chart, 2022). RACI matrix defines the different roles as responsible, accountable, consulted, and informed. One critical part of information system design is to follow an architectural approach first presented by Zachman (1987) with specific aspects developed for the web-based systems presented by Molnár and Tarcsi (2013). We adapt these approaches starting from the operational view through the system view, down to the technical view for the web-based GMIS for ECHO with a focus on information sharing and the need for migration from the ECHO Project Information System to ECHO Collaborative Network Organization GMIS.

2.3 Information Sharing

A key part of any organization's data management is information sharing, without effective data sharing operational capacity and decision-making may be disturbed. The system used for information sharing should be compatible with the whole organization, otherwise, it can lead to unnecessary transcribing and rekeying of the information when it is transferred between systems. Other important requirements for information-sharing systems are ease of access and fast data search. If the needed information is not readily accessible, it can cause frustration and reduce organizational effectiveness. (Gordon, 2013)

Information sharing can be divided into two separate categories, internal and external. Interactions in the work community can lead to a higher level of cooperation and better work performance (Peters & Manz, 2007). Effective internal communication increases inspiration, engagement, and productivity, but it is not easy to find the correct internal communication system for the organization and its culture. Newsletters and emails are not effective communication. It is important that the internal channels are immediate, fast, targeted, measurable—and mobile. An organisation's internal communications can for example include different mailing lists, where each department (ECHO CNO entity) has its own list. For the purposes of internal information sharing, an efficient tool is an organization-wide intranet, where all needed information is easily accessible when needed. With external information sharing, one commonly used method is the organization's web page. Any news or other shared information can be easily accessed by all visitors through a www-site. Building a specific extranet for the organization's various partners can also be beneficial, so they can easily communicate without having access to the intranet. Smooth communication channels are extremely useful in building stakeholder engagement and ensuring effective end-user collaboration (Ruoslahti & Tikanmäki, 2019).

Sharing cybersecurity information between different organizations can be very helpful. It can improve response to threats, help to defend against potential attackers, and mitigate damages. Information sharing can also help to improve relations and build trust between organizations. Information sharing is not without any challenges, legal issues may arise when different countries have different definitions of protecting classified information. Therefore, it is very important to use the right information-sharing models and frameworks to achieve efficient data sharing between organizations (Rajamäki & Katos, 2020).

2.4 Information System Migration

As information systems grow older, they can become outdated and may no longer serve their intended function. Thus it is often very important for an organization to migrate to a more evolved and modernized platform that can provide all the needed services. Information System Migration projects can be very complex, as they should be completed without any loss of data or functionalities, and access to the system should be available during the migration process (Klettke & Thalheim, 2011). Some factors that may further increase the complexity are dealing with legacy information systems (Limaj, et al., 2020) and migrating on-premises data to a cloud platform (Alharthi, 2023).

For IS migration Klettke and Thalheim (2011) present three different strategies, each with its own strengths and weaknesses. Choosing the right strategy to use depends on system complexity, modularization of the legacy system, frequency of the data changes, amount of data in the database, and accepted delay time of the system.

- Big bang – In this strategy, a completely new system is developed at once. During the development process, the legacy system serves as the operating system. Although the migration process can be easy to manage, development times are usually long, and there is no system availability during the transition.

- Chicken little – In this strategy, the system is modularized, and all components are migrated separately. System availability is high through the transition and the individual development and migration times are short. This strategy requires two parallel systems to exist at the same time, which can prove problematic when they must interact with each other.
- Butterfly – Combination of both previous strategies. In this strategy, the focus is on the migration of the database. The legacy system is designated as read-only storage and all the required changes are done to the new system. This mitigates the risk of possible data loss, and it provides a lot of already available data with which to test the new system. Although this strategy can lead to long development times and the design of the old system has a big impact on the new one.

3. Environment and the Relevance Cycles

The relevance cycle initiates design science research with an application context that not only provides the research with requirements as inputs but also defines the acceptance criteria for the final evaluation of the research results (Hevner & Chatterjee, 2010). In this study, the application domain deals with the ECHO project and its information systems, the planned ECHO collaborative network organisation (CNO), and the requirements for the GMIS coming from the environment.

3.1 The ECHO Project

European network of Cybersecurity centres and competence Hub for innovation and Operations (ECHO) is a European Union project founded under European Commission's Horizon 2020 Program. The aim of ECHO is to develop a unified cyber defence ecosystem for the European Union and improve the secure cooperation of its members. This is achieved through multiple different ECHO concepts, such as ECHO Governance model for managing current and future partners, ECHO Federated Cyber Range for cyber simulation training, and ECHO Early Warning System for secure sharing of cyber threat information. One of the goals of ECHO is to raise EU citizens' cybersecurity awareness, helping everyone to be more informed about potential threats. This education is also expanded into different industry fields to help them mitigate possible attacks. Creating an EU-spanning network of cybersecurity hubs is the main challenge of the project, which it aims to complete in the near future. (ECHO, 2022)

The information system of the ECHO project consisted of the following parts:

1. Websites that share public deliverables, scientific publications, newsletters, and other information intended for general distribution. They also included a service related to supporting partnership engagement.
2. Internal SharePoint, Teams, and One Drive for Business platform which included project repositories for contract documents, changes, review-related documentation, reporting documentation, contact information, templates, work documents for deliverables, final versions of all deliverables, internal work documents, agendas, minutes, etc. The project's telephone conference system was also implemented using this.
3. Zenodo open repository used for open datasets.

3.2 Transitioning from project to network

One of ECHO's big milestones is the transition from the initial ECHO project into the complete ECHO CNO. This transition can be summarized as moving from the Current Operating Model (COM) to the Target Operating Model (TOM) (Mladenova, et al., 2022). The final goal is to build CNO that have Central Hubs, National Hubs, and Service Groups (Penchev, 2021). Central Hub is ECHO CNO's overall management (including the governing level of boards/councils) body. National hubs are established on a national level and serve as contact points between the Central Hub and national authorities and organizations. Service groups are formed around the delivery of specific services, such as the ECHO Federated Cyber Range (E-FCR) or the ECHO Early Warning System (E-EWS). The ECHO GMIS is mainly intended for internal use within the organization. The Catalogue (Market Place) is intended for service providers (Service Groups) and their customers, and the web pages serve all stakeholders. (Penchev, G. et al., 2021, p. 45):

To ensure sustainability after the end of the ECHO project, the need for a transition plan was implemented. As project partners have varying levels of commitment, capabilities, resources and, experience, the plan must take into account the capacity of individual organizations to change. Figure 1 shows the approach used in transition

planning and change management, where Initial Operating Capability (IOC) is achieved in the first phase, then plans are implemented in the second phase and the CNO can move to Final Operating Capability (FOC). The goal is to reach the IOC by the end of the ECHO project in February 2023 (Mladenova, et al., 2022).

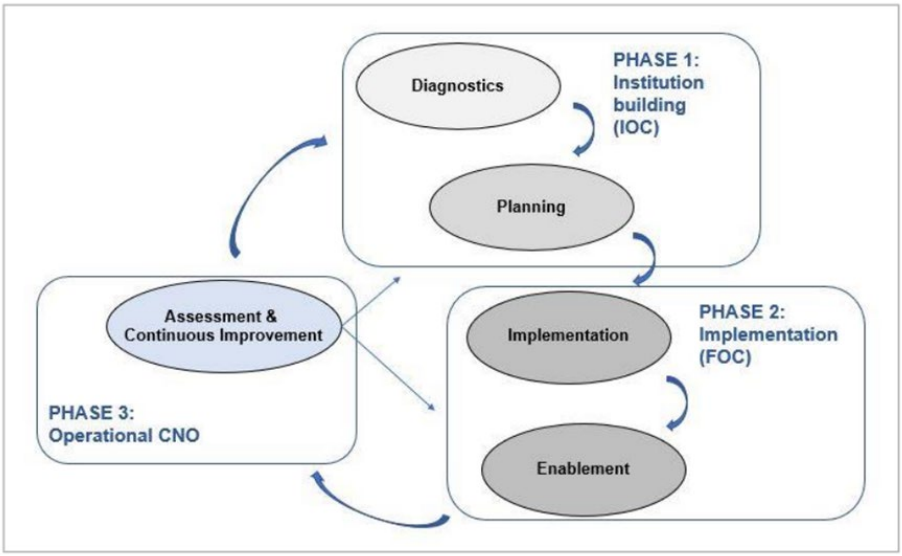


Figure 1: Approach to transition planning and managing change (Mladenova et al. 2022, 134)

3.3 ECHO GMIS Requirements

Deliverable ECHO D3.13 defines the high-level requirements for GMIS: (1) support processes, (2) ensure efficient information sharing, and (3) help to meet the requirements of laws and regulations. The ECHO CNO's core processes are strategic planning, partnership development, innovation management, and catalogue and customer relations, and other processes are HR management, financial management, data management including document handling, portfolio management, continuity management, and external relations management (Penchev, et al., 2021).

Communication between different partners is a key factor in the governance of collaborative networked organizations. The ECHO GMIS should solve this by having a communication platform that is community-based and accessible to all different ECHO members and partners. The platform will include, for example, an intranet, extranet, web pages, calendar management, and document repositories.

The ECHO CNO provides services through an electronic marketplace, which is a digital service that allows individuals or merchants to enter into sales or service contracts with merchants. This means that the Directive on Security of Network and Information Systems (NIS/NIS 2 Directive) applies to GMIS. The General Data Protection Regulation (GDPR) applies to all personal data. This means that GMISs must define controllers and processors who demonstrate compliance with its requirements with specific documents, including relevant logs, policies, and procedures. In addition, the national legislation of the countries where ECHO CNO operates must also be observed.

4. Design Cycle

It has been decided to implement GMIS with Microsoft SharePoint. Next, with the help of blogs found on the internet, etc., we will investigate what the future platform should be like, how internal communication should be implemented, how to ensure compliance with laws and regulations, and how to carry out self-evaluation.

4.1 Platform planning

Microsoft (2022) describes SharePoint as a portal, where you can engage, learn, and collaborate. It consists of portals, sites, and pages. SharePoint as an internal communications tool has many possibilities and solutions. It is part of Microsoft 365 and is possible to combine it with Teams for chats and meetings, Yammer with live events, Outlook for emails, Stream for channels to follow, and Viva for connecting. You can have a portal with

news, connecting, events, sharing, announcements, networking, communicating, training, and learning, employee onboarding, measuring of engagement, good user experience and much more.

According to Microsoft, an intelligent intranet is made of rapid deployment, has been built with user-experience in mind, has a personalized experience, and has software as a service, which means that it is updated automatically and always has the latest technology. Modernizing the intranet is a big job that needs resources but will be worthwhile because it makes the intranet dynamic and scalable, it can target the audience and works on any device, there is central governance of permissions, security, and sharing grows the adaption due to relevant content, user-friendliness, and responsiveness. A modern intranet is where you start your day, it promotes collaboration and has everything in one place. SharePoint has an intuitive discussion board that can be used by the employees to discuss work subjects or give opinions. It is fast to create and can be embedded on the intranet or on the departmental home pages. Discussions can also be highlighted so they are shown under the featured view (Jones, 2014).

There are Wiki pages, where a document library can be edited or added to by anyone. Existing pages can be linked together, and it is possible to link to new pages or create pages. You can use Announcement lists for announcements that expire on a specific date and are shown on the intranet landing page or any other page. You can search in it and add alerts to show when a new announcement is made. If you need an opinion, you can use a survey for input. They can come as a pop-up, be exported to Excel, and get stored as SharePoint list. It is an easy way to gather data through information and replies. There is an alert possibility for changes in contents, a community site template for social networking around a given subject with gamification giving points for completing tasks, discussion forums and an activity view of the members, you can add a web part to wiki or content pages for video on the intranet with a thumbnail and create a blog site template where you can add and manage your posts, comments, and categories. You can also use Yammer as social networking almost like Facebook.

There are add-ons for DocRead for tracking and distributing tasks and DocSurvey for mandatory quizzes. And you can use Delve for monitoring usage inside SharePoint (Jones, 2014). The content tagging. It is almost like Hash tagging used in Twitter but is a Tag cloud that is larger when it is a popular tag and it can be clicked to see the tagged content, to follow a tag, to add notes, and to view people who are following the tags. SharePoint also has a mobile-friendly version that works on SharePoint Online and Microsoft has created a mobile app for iOS and Android.

SharePoint Team site can be used for communication and collaboration, like for projects, tracking status, planning events, etc. Or you can use Communications site for communicating messages to a wide audience, where there are a few content creators and a wide audience for content like HR-related, guidelines, policies, micro-sites, etc. Or a hub site that shows what is happening across sites, is searchable, and creates cohesion with shared navigation, look and feel (Shareknowledge, 2020). Even though there seem to be many ways to use SharePoint as an internal communications tool there are opinions about it not being good enough for that purpose. SharePoint is lacking user engagement when it is used for enterprise communications being a document management and storage system promoting collaboration, but it is not a communications platform and cannot compete with modern employee experience platforms (Unily, 2022). They reference to a Forrester survey of business leaders using SharePoint from 2013 where 64% of business leaders could not see the expected degree of adoption, 62% meant it was difficult to use, 49% preferred other tools, 44% could not see any added benefits and 44% meant that it did not reach the functional requirements. SharePoint is a solution that is often chosen by IT because it is an easy choice in a Microsoft-led organization and will integrate with the existing Microsoft solutions they already have. Unily (2022) lists five reasons why SharePoint doesn't cut it for internal communication:

1. SharePoint is lacking essential communications functionalities such as engaging, meeting expectations, personalisation, availability on any devices, and delivery through many channels and in many formats.
2. SharePoint cannot unite communication channels such as integrating to other than Microsoft solutions. One example is Slack and meeting employees on channels they use daily.
3. SharePoint means relying on IT resources, where all changes need to be done by IT.
4. The lack of measurement and analytics capabilities; there will be no possibility to see what employees are engaging with, and when and how they are doing it.
5. It fails for frontline workers as they do not have access to computers, and it is not working optimally on mobile.

Harris (2022) gives a broader view of why SharePoint is chosen in organizations and if it is a good choice. SharePoint is a natural choice for organizations that are using Microsoft 365 even though it doesn't always meet managers' needs as they are today. The benefits of SharePoint are document management, centralizing policies and procedures, collaborations, sharing, storing, and organizing information. It is natively integrated with Outlook, Teams, Yammer, and OneDrive. It allows to access information from anywhere through SharePoint online and offers enhanced security with access management and encryption. It has a good content management system and approved content can be published on the internet and social media platforms. There are possibilities to automate and streamline processes and business operations, and it can be integrated with applications, email programs, and browsers giving a good user experience. The negative side of SharePoint is that it is mostly for top-down communications, it could be integrated with Teams for better employee engagement and collaboration in decision-making but is not designed for two-way internal communication. SharePoint is not good for frontline workers due to the need for Microsoft 365 license. Information overload is one issue as it is difficult to send the correct information to the correct people unless they are using targeting, which still can be improved in SharePoint. SharePoint also needs a third-party solution if branding is important as SharePoint alone will look like Microsoft intranet. Also, the engaging features are limited and to meet employee expectations third-party solutions with more features like employee advocacy, desk booking, and more is recommended (Harris, 2022).

4.2 Internal communications

The intranet needs to be well-designed, user-friendly, responsive, and full of engaging information which the users require (Pocketstop, 2018). Internal communication best practices to give the employees the feeling of importance and purpose, remove the hierarchical information flow, and avoid gossip and rumours include (Pocketstop, 2018):

1. Internal communications strategy – think, strategize, and Plan
2. Use the right internal communications tools
3. Use visual communication – screens, SOME, motivational quotes, goals, accomplishes
4. Provide channels for feedback and ideas
5. Encourage to give feedback
6. Avoid communication overload – brief, to the point, relevant, automation of key messages
7. Promote employee resources and training
8. Encourage cross-departmental communication and collaboration

According to Temkin and Lucas, the five I's that drive employee engagement are to inform, inspire, instruct, involve, and incent (Temkin & Lucas, 2018). These can also be used as guidance for internal communication. Strachan (Strachan, 2022) lists six best practices to build a successful intranet and recommends considering the three E's: engagement, enablement, and empowerment. He also says that intranet should have relevant and encouraging information. The best practices include defining intranet with strategic goals, where you can measure the KPIs for Employee adoption through how many are using it, time to access knowledge through the time from search to result, average click rank through search engine effectivity and employee satisfaction through feedback surveys. The second part is about consulting the employees to find out what is most important about the corporate intranet, what they like/ do not like of the intranet, what information they want to/do not want to see, what user experiences would increase the usage of intranet. The third part is to unify the content across the company. The fourth part is to personalize the experience through using location, departments, roles, tenures, projects, product lines and markets. The fifth part is about having a balance between business and personalized content. And the last is about improving the employee experience.

4.3 Compliance with Laws and Regulations

As ECHO's operating environment is the European Union, there are couple of directives we have to be compliant with. The first one is the General Data Protection Regulation (GDPR), which requires all organizations to have rigorous data handling policies with data controllers and processors. The other one is the Network and Information Security (NIS) Directive, which is an EU-wide legislation on cybersecurity, with a specific aim to achieve a high common level of cybersecurity across the Member States. This Directive applies to ECHO GMIS, as the platform will have an online marketplace for the services offered by various partners.

The following sections of the NIS directive need to be addressed when designing the future GMIS (ENISA, 2022):

- Human resource security – security training programs are required for employees with CIS related responsibilities.
- Information system security risk analysis – during the designing process, regular risk analysis are conducted so possible risks are assessed as early as possible.
- Information system security audit – taking account the regularly updated risk analysis, all critical assets are frequently audited to ensure compliance with regulations.
- Ecosystem mapping – during the design process, all documentation of GMIS ecosystem is frequently updated.
- Information system security policy – building up to the risk analysis, an information system security policy will be established and maintained.
- Ecosystem relations – the interfaces between ECHO GMIS and third parties are designed so that potential risks are mitigated.

As the EU is a very diverse area, there are also various national laws and regulations concerning data privacy and handling. All of this has to be taken into consideration when building the ECHO GMIS.

4.4 Self-evaluation

During the ECHO project, there have been several workshops (e.g., the ECHO WP3 Workshop, 10-11 Nov. 2022) from which feedback has been received for further development of the project. As the future GMIS will have multiple national and international shareholders, this feedback has been very beneficial for the clarifications and confirmations it has provided. Self-evaluation and gathering feedback will continue until the end of the project.

5. Implementation plan

The overall migration is a key effort to move from ECHO Project to ECHO CNO. A big challenge is to figure out how to transfer all ECHO Project materials to the new GMIS or other repositories. The transferred material can be any digital material, such as a document, image, video, folder, or entire folder structure with documents. Once the material to be transferred is defined, a digital object is formed from it. In this context, it is given a persistent ID and the desired metadata is defined. After this, the transfer to the desired new location is performed.

The first step was to design a new folder structure for the old SharePoint. One big change between ECHO Project's SharePoint portal and the future GMIS is the access and user rights. In the project's portal all users had access to everything, but as CNO partners are not the same as ECHO project partners, this should be changed. Even after the project, the project partners have to be guaranteed access rights to the project outputs. As some of the project folders contain confidential information, not all members of the future CNO can be automatically given access to them. Folder access and user rights should be determined using the RACI matrix.

Creating and implementing a new information system is a major task that must be planned carefully and in sufficient time. With the ECHO project ending, there is no time to build a completely new system before the CNO starts operating, so we have to make do with the old system at first. The main thing is that the system starts small and grows with the Central Hub. GMIS and web visibility should be outsourced and in the first phase, the continuation of their maintenance is agreed upon year by year for a maximum of 5 years with the current administrators. A lot of additional work will be required during the transition, as CNO partners are not the same as the project partners and may have different levels of access and user rights. Once the CNO's operations are established, the GMIS planning project can be initiated. The size of this information systems development project depends on the scope and scale of the CNO's operations.

6. Discussion and Conclusions

The DYNAMO project (10/2022-9/2025) creates tools based on the refinement of the existing solutions of the ECHO project for a cyber situational picture to support decision-making. Information sharing allows organisations to leverage the collective knowledge, experience, and analytical capabilities of their sharing partners in a community of interest, improving the cybersecurity of multiple organisations. Towards this goal, the tools developed in DYNAMO convert heterogeneous data into common formats and protocols, which enables automation and allows organizations to quickly exchange information. Privacy preservation techniques are used to protect sensitive and classified information without preventing access and sharing of information. This paper examines the design of the ECHO governance and management information system and describes

the structure of such an information system, which best supports the organisational processes and information-sharing needs of the collaborative network.

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

Alharthi, D., 2023. Secure Cloud Migration Strategy (SCMS): A Safe Journey to the Cloud. *Vol. 18 No. 1 (2023): Proceedings of the 18th International Conference on Cyber Warfare and Security*, pp. 1-6. DOI: https://doi.org/10.34190/iccws.18.1.1038 [Accessed 10 March 2023].

CONCORDIA, 2022. *CONCORDIA - Home*. [Online] Available at: https://www.concordia-h2020.eu/ [Accessed 13 December 2022].

CyberSec4Europe, 2022. *Cyber Security for Europe - Home*. [Online] Available at: https://cybersec4europe.eu/ [Accessed 13 Dec. 2022].

DYNAMO, 2023. *DYNAMO Project - Home*. [Online] Available at: https://horizon-dynamo.eu/ [Accessed 18 January 2023].

ECHO, 2022. *ECHO - Home*. [Online] Available at: https://echonetwork.eu/ [Accessed 13 December 2022].

ENISA, 2022. *Minimum Security Measures for Operators of Essentials Services*. [Online] Available at: https://www.enisa.europa.eu/topics/cybersecurity-policy/nis-directive-new/minimum-security-measures-for-operators-of-essentials-services [Accessed 19 January 2023].

Gordon, K., 2013. *Principles of Data Management - Facilitating Information Sharing*. Second Edition. Swindon: BCS Learning & Development Limited.

Harris, Y., 2022. *Is SharePoint a good internal communication tool?.* [Online] Available at: https://powell-software.com/resources/blog/sharepoint-internal-communication-tool/ [Accessed 19 January 2023].

Hevner, A. & Chatterjee, S., 2010. *Design research in information systems: theory and practice*. New York: Springer Science and Business Media.

Hevner, A., March, S., Park, J. & Ram, S., 2004. Design science in information systems research. *MIS Quarterly*, 28(1), pp. 75–105.

Jones, M., 2014. *11 Ways to Use SharePoint for Internal Communications*. [Online] Available at: https://community.aiim.org/blogs/mark-jones/2014/08/21/11-ways-to-use-sharepoint-for-internal-communications [Accessed 18 January 2023].

Klettke, M. & Thalheim, B., 2011. Evolution and Migration of Information Systems. In: Embley, D., Thalheim, B. (eds) *Handbook of Conceptual Modeling*. Heidelberg: Springer, pp. 381–419.

Limaj, E., Bernroider, E. & Ivanova, M., 2020. Facing Legacy Information System Modernization in Scaling Agility in the Banking Industry: Preliminary Insights on and Non-technical Barriers. In: Joey George, Paul, Rahul De' (Ed.), *Proceedings of the 41st International Conference on Information Systems (ICIS)*. AIS Association for Information Systems, pp. 1–9.

March, T. & Smith, G., 1995. Design and natural science research on information technology. *Decis Supp*, 15(4), pp. 251-266.

Mirosoft, 2022. *Step 2: Review Microsoft 365 communication tools*. [Online] Available at: https://learn.microsoft.com/en-us/sharepoint/review-communication-apps?source=recommendations [Accessed 19 January 2023].

Mladenova, I., Shalamanov, V. & Shalamanova-Filipova, A., 2022. Governance Consulting Services and Tools: Transition Planning and Implementation for Collaborative Networked Organisations in the Cyber Domain. *Information & Security: An International Journal*, 53(1), pp. 131-146.

Molnár, B. & Tarcsi, Á., 2013. Design and architectural issues of contemporary web-based information systems. *Mediterr. J. Comput. Netw*., Volume 9, pp. 20-28.

Penchev, G. et al., 2021. *ECHO D3.3 GOVERNANCE MODEL DESCRIPTION*. [Online] Available at: https://echonetwork.eu/wp-content/uploads/2021/03/ECHO_D3.3.Governance-Model-Description.pdf [Accessed 18 November 2022].

Penchev, G., 2021. Planning and Implementing Change in Cyber Security Network Organisations. *Information & Security: An International Journal*, 50(1), pp. 89-101.

Peters, L. & Manz, C. M., 2007. Identifying antecedents of virtual team collaboration. *Team Performance Management: An International Journal*, 13(3/4), pp. 117-129.

Pocketstop, 2018. *Rethinking Internal Communications*. [online] June 15. [Online] Available at: https://pocketstop.com/blog/rethinking-internal-communications/ [Accessed 19 January 2023].

Pocketstop, 2018. *What Is an Internal Communication System?* [Online] Available at: https://pocketstop.com/blog/what-is-an-internal-communication-system/ [Accessed 19 January 2023].

RACI Chart, 2022. *RACI Chart*. [Online] Available at: https://racichart.org/ [Accessed 19 November 2022].

Rajamäki, J. & Katos, V., 2020. Information Sharing Models for Early Warning Systems of Cybersecurity Intelligence. *Information & Security: An International Journal*, 46(2), pp. 198-214.

Ruoslahti, H. & Tikanmäki, I., 2019. Complex Authority Network Interactions in the Common Information Sharing Environment. In: Bernardino, Jorge; Salgado, Ana; Filipe, Joaquim (Eds.) *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2019).* Setúbal: Science and Technology Publications, pp. 159 – 166.

Shareknowledge, 2020. *3 new ways SharePoint impacts communication in the workplace*. [Online] Available at: https://www.shareknowledge.com/blog/3-new-ways-sharepoint-impacts-communication-in-the-workplace [Accessed 19 January 2023].

Smallwood, R. F., 2019. *Information Governance*, 2nd Edition. John Wiley & Sons, Inc.

SPARTA, 2022. *SPARTA - Home*. [Online] Available at: https://www.sparta.eu/ [Accessed 13 December 2022].

Strachan, C., 2022. *6 Intranet Best Practices for a More Engaged Workplace in 2022*. [Online] Available at: https://www.coveo.com/blog/intranet-best-practices/ [Accessed 19 January 2023].

Temkin, B. & Lucas, A., 2018. *Employee Engagement Competency & Maturity Insight Report 2018.* [Online] Available at: https://www.xminstitute.com/research/2018-employee-engagement-maturity/ [Accessed 19 January 2023].

The Chartered Governance Institute UK & Ireland, 2022. *What is corporate governance?* [Online] Available at: https://www.cgi.org.uk/about-us/policy/what-is-corporate-governance [Accessed 18 November 2022].

Unily, 2022. *Why SharePoint doesn't cut it for internal comms*. [Online] Available at: https://www.unily.com/insights/blogs/why-sharepoint-doesnt-cut-it-for-internal-comms [Accessed 19 January 2023].

Yang, T. & Maxwell, T., 2011. Information-sharing in public organizations: A literature review of interpersonal, intra-organizational and inter-organizational success factors. *Government Information Quarterly*, 28(2), pp. 164-175.

Zachman, J. A., 1987. A framework for information systems architecture. *IBM Systems Journal*, 26(3), pp. 276-292.

Students' Application of the MITRE ATT&CK® Framework via a real-time Cybersecurity Exercise

Aunshul Rege [1], Jamie Williams [2], Rachel Bleiman [1] and Katorah Williams [1]
[1] Criminal Justice, Temple University, Philadelphia, USA
[2] The MITRE Corporation, USA

rege@temple.edu
jcwilliams@mitre.org
Rachel.bleiman@temple.edu
Katorah.williams@temple.edu

Abstract: The MITRE ATT&CK framework enables practitioners to understand and track cyber adversary behaviors. Concepts such as social engineering (SE) are not directly captured in current version of ATT&CK as an individual technique, though the application of SE is relevant to many technical behaviors. Utilizing the ATT&CK framework in an educational setting, specifically within a competition focused on SE, allows students to explore adversarial behavior through experiential learning and understand how SE is relevant within cybersecurity. The structure of the framework allows students to see and describe each behavior from the perspective of the adversary, motivating them to compile and question "why" and "how" each individual action contributes to the operational objectives. This paper shares students' mappings of the ATT&CK framework to playbooks they developed during a real-time SE penetration testing competition. Students were given numerous flags to pursue during the competition and this paper will share their playbooks and mappings to the ATT&CK framework. This paper demonstrates that while someone with more knowledge and experience using the framework may map a SE case study differently than multidisciplinary students who are experiencing it for the first time, there is not a single correct way to map onto the matrix. Having students experience this mapping process allows them to understand the breakdown of an adversary's behavior and interpret key tactics and techniques in a way that fits their mapping needs. This paper also demonstrates how a SE case study can be mapped onto the ATT&CK framework despite SE not being the focus of the framework, and that SE uses tactics and techniques that are also relevant to technical cyberattacks. The authors hope to encourage more interdisciplinary cybersecurity education by sharing this experiential learning event.

Keywords: social engineering; pentesting; education; ATT&CK; attack playbooks

1. Introduction

A penetration test is a simulated cyberattack that is authorized by a company or organization to evaluate the security of the company's system. Traditionally, penetration tests are often technical in nature, utilizing such skills as network and application security, scripting, various operating system environments, cryptography, and remote access technologies, among a multitude of others. However, there also exist penetration tests that instead, or in part, rely on social engineering (SE) to evaluate the security of a company's system.

Social engineering is a technique where psychological persuasion of humans is used to "conduct reconnaissance (identify systems operating at target facilities), obtain information intended to secure electronic systems (passwords), or to encourage targets to inadvertently provide access to electronic systems and information (downloading and executing malicious files that are disguised as familiar or benign)" (Rege & Bleiman 2022). Because humans are often exploited in cyberattacks, a social engineering pentest can show how technical security measures can be bypassed without any technical expertise. Further, without the reliance on technical skills, social engineering penetration tests offer students and professionals from multiple disciplines an avenue into cybersecurity.

To promote interdisciplinary engagement in cybersecurity, the authors designed and implemented a social engineering pentesting competition in the summer of 2021. During this competition teams of students used SE to capture a series of flags, created adversarial playbooks, and mapped their playbooks to the ATT&CK framework. The framework, discussed in greater detail in the next section, organizes and describes cyber adversary behaviors and serves as a resource for threat mapping.

The following section describes the ATT&CK framework, including its tactics, techniques, and procedures as well as its various uses in education and competition spaces. Next, the paper describes the 2021 social engineering pen test competition, outlining the various flags that students were tasked with capturing. Then, the results include a deeper dive into five specific flags, discussing each's use regarding the flag and comparing its application across the different teams. The paper ends with a discussion of how ATT&CK applies to social

engineering, key takeaways from student mappings, implications for mitigation, and a consideration of the value of mapping exercises to students.

2. ATT&CK Framework

The MITRE ATT&CK framework enables practitioners to understand and track cyber adversary behaviors. The globally-accessible knowledge base (attack.mitre.org/matrices/enterprise) is maintained by the MITRE Corporation as a means of organizing and analyzing the tactics, techniques, and procedures (TTPs) used by real adversaries. The content within ATT&CK is informed by real-world observations and is continuously maintained via analysis of publicly available cyber threat intelligence (CTI) as well as contributions from the community of ATT&CK users.

ATT&CK is appropriately structured around organizing adversary behaviors into TTPs. Specifically, for each included behavior ATT&CK captures and contextualizes the adversary's:

1. Tactic(s), or "why" the behavior was performed
2. Technique, or "how" the adversary attempted to achieve their tactical goal by performing the behavior. ATT&CK also includes sub-techniques, which are functional equivalents to techniques but describe the specific behavior at a lower level than a technique (where applicable).
3. Procedure, or "what" the adversary specifically did to implement the technique

The organization provided by ATT&CK not only captures the adversary perspective of malicious cyber operations, but also facilitates the creation of a common, shared language for tracking these behaviors. This enables users to apply ATT&CK towards many operational use cases, including (but not limited to):

- Tracking and organizing new as well as known behaviors observed in threat intelligence/case studies
- Developing and aligning defensive countermeasures for specific behaviors
- Prioritizing and communicating behaviors used by a red or other form of offensive assessment team
- Engineering and documenting an organization's current defensive posture, including strengths and potential gaps relative to specific adversary behaviors

Though ATT&CK is selectively scoped to cyber activities (i.e., those directly involving victimized systems modeled into the framework as platforms), the structure and way of organizing TTPs is applicable to wider domains. This structure -- specifically modeling behaviors into why (tactics), how (techniques), and what (procedures) -- is conducive towards modeling and connecting behaviors to defensive countermeasures. Researchers have explored and used the blueprint of ATT&CK to create similar ATT&CK-like representations of differing adversary behaviors.

Concepts such as social engineering are not directly captured in current versions of ATT&CK as an individual technique/object, though the application of social engineering is relevant to many technical behaviors. For example, T1566 Phishing as well as many other behaviors exist where the "human-element" associated with the targeted system is exaggerated. These techniques specifically incorporate or even rely-on elements of persuasion, manipulation, elicitation, and impersonation for successful execution.

2.1 Uses of ATT&CK Framework

The ATT&CK framework has been used to introduce and expose various learners to security concepts, specifically the practice of modeling the cyber behaviors of threats (mitrecorp 2018). ATT&CK provides a structured process and prototype for tracking and understanding malicious activity within the context of analyzing why and how an adversary makes certain decisions and takes corresponding actions. An "ATT&CK-like" approach to modeling threats also lends towards exploring many different domains as well as use cases regarding defensive operations (Nickels et al 2019). The context provided by ATT&CK can also provide focus and a storyline to accommodate competition participants (Spunk 2019). For instance, groups of students mapped a social engineering case study onto the MITRE ATT&CK framework to understand the adversarial mindset for a course project. Having students experience this mapping project allowed them to understand the breakdown of an adversary's behavior and interpret key tactics and techniques in a way that fits their mapping needs (Bleiman et al. 2022).

3. Cybersecurity Competition

The ATT&CK framework was used in the Social Engineering (SE) Penetration Test Competition (SE-PTC) in the summer of 2021. The context of the SE-PTC was to conduct a SE pen test of the cybersecurity lab run by the

authors. Student teams would pose as pen testing firms the authors hired to test their (lab employees') vulnerability to SE attacks (Rege & Bleiman 2022). A total of 16 teams engaged in the competition: 1 high school team, 9 undergraduate teams, and 6 graduate teams (Rege & Bleiman 2022).

Each team had to accomplish a set of predetermined flags (tasks), such as acquiring information about lab employees' mail client and browser, obtaining copies of datasets and publications, convincing lab employees to change information on the lab's website, trying to get hired, and attempting to develop collaborations. Students had to use three SE tactics: OSINT, phishing, and vishing to pursue these flags.

OSINT or Open Source INTelligence involves gathering information from publicly available sources (Nickels et al 2019). Phishing occurs when a target is contacted via email by a cybercriminal who poses as a legitimate organization to lure the target into providing sensitive data, such as intellectual property, user accounts and passwords, or obtain confidential information (Nickels et al 2019). Vishing, or voice solicitation, has the same objectives as phishing (acquire sensitive information), but uses the phone or other voice over IP technologies (Rege & Bleiman 2022). More information about the competition structure and logistics is available in (Rege & Bleiman 2022).

Students then had to develop corresponding playbooks that documented their various strategies for each flag (Rege & Bleiman 2022). More specifically, students had to map their playbook strategies to the ATT&CK framework. For a more thorough discussion of the competition structure, logistics, and student experiences, please see the Rege & Bleiman (2022) paper.

3.1 Flags

Students were given numerous flags to pursue during the competition, which are displayed in Table 1 below. This paper will focus on flags 2, 3, 6, 10, and 12 to show student playbooks and mappings to the ATT&CK framework (marked with * in Table 1); these details for the remaining flags can be found in Appendix 1.

Table 1: Flags for the 2021 Summer SE Pen Test Competition

Flag #	Flag Description	# pursuing teams
1	Get the CARE Lab's Twitter account to Like a tweet posted by a competing team	10; 4
2*	Get the CARE Lab's Twitter account to Retweet a tweet posted by a competing team	8; 5
3*	Get the CARE Lab to disclose its office location and/or phone number	14; 5
4	Get the CARE Lab to disclose its mail client	7; 5
5	Get the CARE Lab to disclose its browser information	8; 4
6*	Get the CARE Lab to share its critical infrastructure ransomware incident dataset	11; 6
7	Get the CARE Lab to share its publications	9; 4
8	Get the CARE Lab to share its conference powerpoints	9; 4
9	Convince the CARE Lab to hire the team to host the CARE Pod	5; 5
10*	Convince the CARE Lab to update the lab website	7; 5
11	Convince the CARE Lab to update a rival team's page on the lab website	12; 5
12*	Collaborate with the CARE Lab to develop education projects	5; 4
13	Collaborate with the CARE Lab to be a guest on the CARE Pod	10; 4

3.1.1 Flag 2: Get the CARE Lab's Twitter account to Retweet a tweet posted by a competing team

For this flag, teams had to convince the CARE Lab employees to 'Retweet' one of their tweets. Overall, eight teams pursued this flag; five of the teams' playbooks are shared here (Figures 1-5).

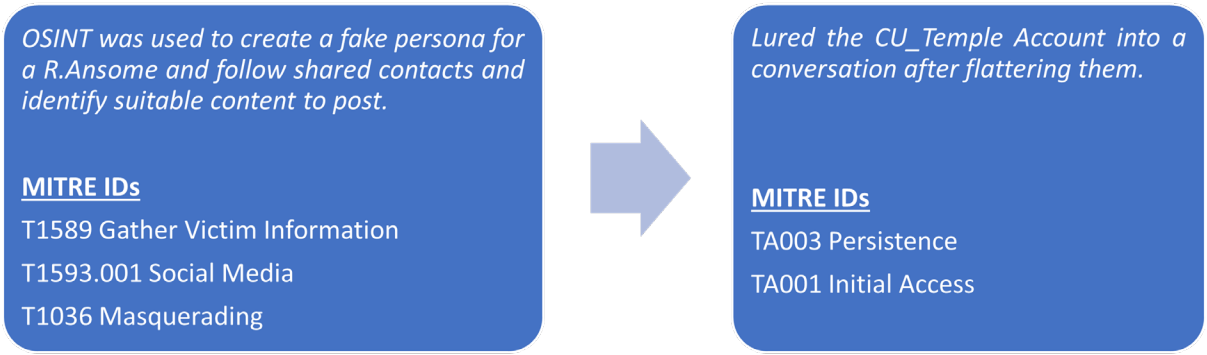


Figure 1: G1 playbook

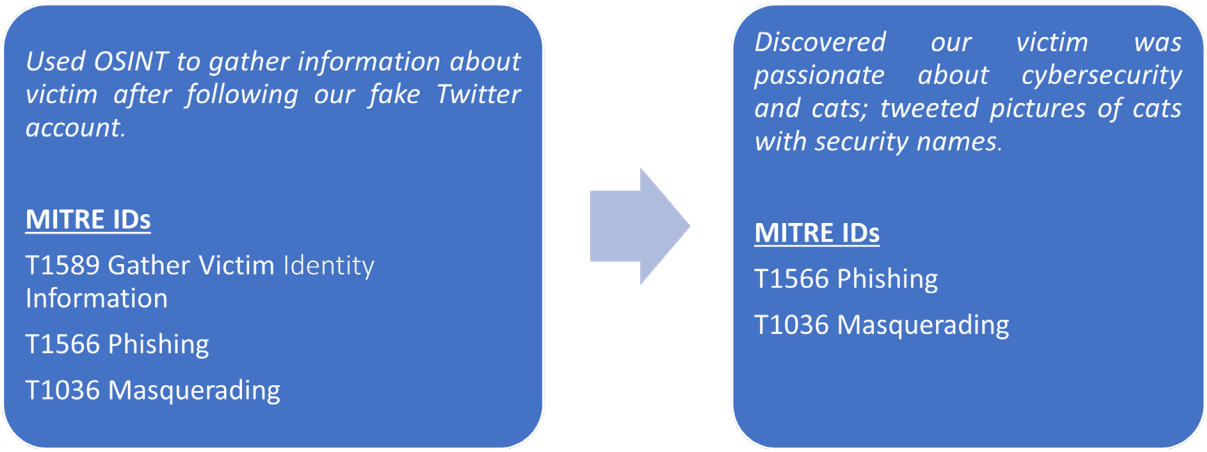


Figure 2: G2 playbook

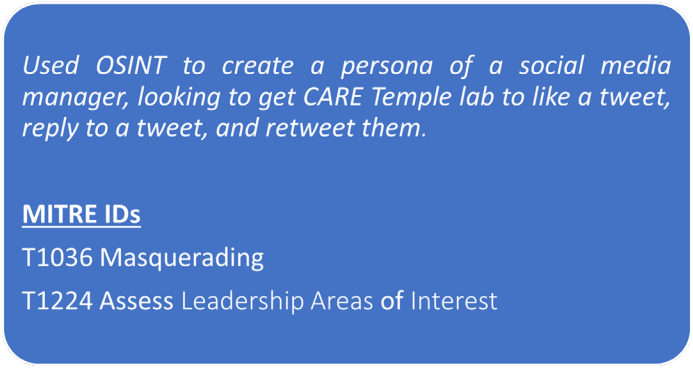


Figure 3: UG2 playbook

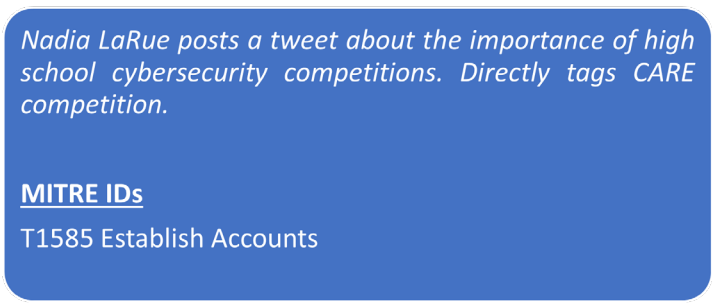


Figure 4: UG3 playbook

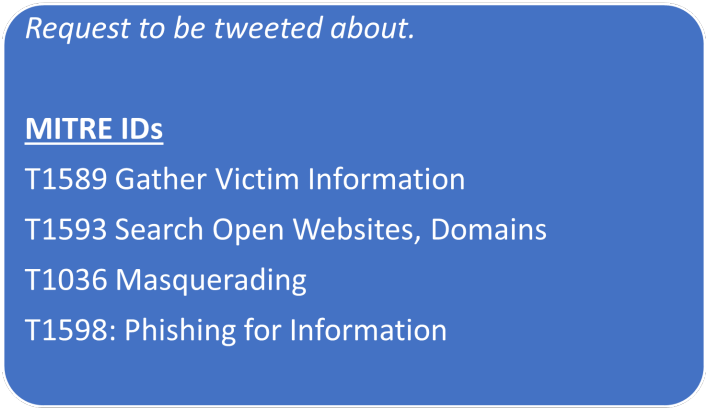


Figure 5: HS playbook

3.1.2 Flag 2 Analysis

The playbooks above indicate that the most commonly used technique was T1036: Masquerading, followed by T1589: Gather Victim Information. These are understandable given the objective; students had to collect OSINT on the Lab employees and correspondingly use that to masquerade and take on an identity. Interestingly, there were some techniques that were only used once across the various teams. Consider the tactics TA0001: Initial Access and TA0003: Persistence. While all teams did engage in both these techniques, only a few actually documented it. Furthermore, one team used a sub-technique, namely T1593.001: Social Media, while other groups primarily used the broader technique T1593: Search Open Websites/Domains. Thus, while many students did use social media in their OSINT, they may have just chosen to use the broader technique instead of the specific one.

Table 2: ATT&CK techniques identified by student teams for Flag 2

Technique ID	Tactic ID	Total	G1	G2	UG2	UG3	HS
T1036: Masquerading	TA0005: Defense Evasion	4	X	X	X		X
T1566: Phishing	TA0001: Initial Access	2		X			X
T1585: Establish Accounts	TA0042: Resource Development	1				X	
T1589: Gather Victim Information	TA0043: Reconnaissance	3	X	X			X
T1593: Search Open Websites/Domains	TA0043: Reconnaissance	2	X				X
N/A	TA0003: Persistence	1	X				

Majority of the playbooks above demonstrate that T1589: Gather Victim Information and T1593: Search Open Websites/Domains, which both belong to TA0043: Reconnaissance, were part of the first stage and thus aligned with the ATT&CK framework. Interestingly, T1036: Masquerading, which is part of TA0005: Defense Evasion (stage 7), is also listed as part of the first stage. The framework defines this technique as "the name or location of an object, legitimate or malicious, is manipulated or abused for the sake of evading defenses and observation. This may include manipulating file metadata, tricking users into misidentifying the file type, and giving legitimate task or service names". However, students may have understood this technique as part of developing pretexts and personas, i.e., 'masquerading' as someone else.

The tactic placement of the techniques highlighted by the students also potentially sheds light on their perceived process for completing this flag. Despite the breadth of their actions (techniques), almost all of the playbooks included tactical decisions based on the shared goals of gathering information (TA0043: Reconnaissance) and/or avoiding detection (TA005: Defense Evasion). Finally, while some students listed TA0003: Persistence, they did not identify a specific framework technique.

3.1.3 Flag 3: Get the CARE Lab to disclose its office location and/or phone number

For this flag, teams had to get information about the CARE Lab, specifically to get employees to share their office location and/or phone numbers. Overall, 14 teams pursued this flag; five of the teams' playbooks are analyzed here.

3.1.4 Flag 3 Analysis

The majority of techniques identified in the playbooks were part of the TA0043: Reconnaissance tactic. A noteworthy observation is the variety of techniques that were used in this tactic: T1589: Gather Victim Information, T1591: Gather Victim Org Information, T1591.001: Determine Physical Location (the actual flag in this case), T1593: Search Open Websites/Domains, T1593.001: Social Media, T1594: Search Victim Owned Websites, and T1598: Phishing for Information. This demonstrates that students utilized an assortment of techniques to find information about the CARE Lab. Interestingly, some techniques such as T1036: Masquerading and T1113: Screen Capture, which appear in later stages of the enterprise matrix, TA0005: Defense Evasion and TA0009: Collection respectively, were used in the initial stages of the playbooks. The possible use of T1036 has been discussed above (see Flag 2 analysis).

Table 3: ATT&CK techniques identified by student teams for Flag 3

Technique ID	Tactic ID	Total	G1	G2	UG2	UG3	HS
T1036: Masquerading	TA0005: Defense Evasion	4	X	X	X		X
T1113: Screen Capture	TA0009: Collection	1					X
T1199: Trusted Relationship	TA0001: Initial Access	1		X			
T1585: Establish Accounts	TA0042: Resource Development	1		X			
T1589: Gather Victim Information	TA0043: Reconnaissance	1					X
T1591: Gather Victim Org Information	TA0043: Reconnaissance	2		X			X
T1591.001: Determine Physical Location	TA0043: Reconnaissance	1					X
T1593: Search Open Websites/Domains	TA0043: Reconnaissance	2			X		X
T1593.001: Social Media	TA0043: Reconnaissance	1	X				
T1594: Search Victim Owned Websites	TA0043: Reconnaissance	1		X			
T1598: Phishing for Information	TA0043: Reconnaissance	2		X			X
N/A	TA0003: Persistence	1			X		

3.1.5 Flag 6: Get the CARE Lab to share its critical infrastructure ransomware incident dataset

For this flag, teams had to convince the CARE Lab employees to 'Retweet' one of their tweets. Overall, 11 teams pursued this flag; five of the teams' playbooks are analyzed here.

3.1.6 Flag 6 Analysis

The playbooks above and Table 4 below show fewer techniques being used for TA0043: Reconnaissance compared to flag 3, which tried to get the Lab employees' office location. This may be related to what students envisioned an adversary would have to achieve in order to complete the flag (i.e., collecting or gathering information from a victim system, particularly the data from a specific website). The specific scope of the flag and resulting limited breadth of observed techniques mapped by the student playbooks can be used as valuable inputs to preemptively model and defend against this type of threat. A real adversary may vary their actions (techniques), yet potentially rely on successful completion of specific identified tactics and techniques central to completing this objective.

However, an interesting observation can be seen in UG3's playbook, which is the only one that captured technique T1005: Data from Local System and thereby captures the very essence of this flag (to exfiltrate the ransomware dataset). However, it is listed as the only technique throughout the playbook, which does not identify any reconnaissance (TA0042) techniques. Another interesting observation is the absence of TA0042: Resource Development, a tactic that was applied in Flags 2 and 3 above.

Table 4: ATT&CK techniques identified by student teams for Flag 6

Technique ID	Tactic ID	Total	G1	G2	UG2	UG3	HS
T1005: Data from Local System	TA0009: Collection	1				X	
T1036: Masquerading	TA0005: Defense Evasion	4	X	X	X		X
T1589: Gather Victim Information	TA0043: Reconnaissance	1					X
T1593: Search Open Websites/Domains	TA0043: Reconnaissance	4	X	X	X		X
T1598: Phishing for Information	TA0043: Reconnaissance	1					X
N/A	TA0003: Persistence	2		X	X		

3.1.7 Flag 10: Convince the CARE Lab to update the lab website

For this flag, teams had to convince the CARE Lab employees to 'Retweet' one of their tweets. Overall, seven teams pursued this flag; four of the teams' playbooks are analyzed here.

3.1.8 Flag 10 Analysis

The objective of this flag was to get the Lab to update its website. The student's mappings suggest a broader interpretation of how (tactics) and what (techniques) could be done to complete this objective. This suggests that the flag provided a wider possibility for adversary creativity, which widens the defenders attack surface while also creating opportunities to identify, track, and potential attribute threats based on their patterns.

In addition to the usual tactics and techniques seen thus far, one interesting tactic/technique used was TA0009: Collection/T1113: Screen Capture. The description for T1113 states that "the next goal after collecting data is to steal (exfiltrate) the data" (and therefore might appear a technique for the last tactic). However, it also states "Collection consists of techniques adversaries may use to gather information and the sources information is collected from that are relevant to following through on the adversary's objectives", which may explain that (in this specific context) students took a screenshot to further their objective for changing the website. Another interesting tactic choice was TA0040: Impact, which is the last tactic in the framework's chain. Interestingly, the technique used here is T1565: Data Manipulation, which aligns with this flag's objective. However, only one instance of this tactic/technique combination was found in the playbooks.

Table 5: ATT&CK techniques identified by student teams for Flag 10

Technique ID	Tactic ID	Total	G1	G2	UG3	HS
T1036: Masquerading	TA0005: Defense Evasion	3	X	X		X
T1113: Screen Capture	TA0009: Collection	1				X
T1199: Trusted Relationship	TA0001: Initial Access	1		X		
T1565: Data Manipulation	TA0040: Impact	1			X	
T1589: Gather Victim Information	TA0043: Reconnaissance	1				X
T1593: Search Open Websites/Domains	TA0043: Reconnaissance	2	X	X		
T1593.001: Social Media	TA0043: Reconnaissance	1				X
T1594: Search Victim Owned Websites	TA0043: Reconnaissance	1	X			
T1598: Phishing for Information	TA0043: Reconnaissance	2	X			X
N/A	TA0003: Persistence	2	X	X		

3.1.9 Flag 12: Collaborate with the CARE Lab to develop education projects

For this flag, teams had to convince the CARE Lab employees to 'Retweet' one of their tweets. Overall, five teams pursued this flag; three of the teams' playbooks are analyzed here.

3.1.10 Flag 12 Analysis

Given the broader nature of the objective of this flag (collaboration), it was interesting that few tactics/techniques were used. Students once again conducted OSINT (TA0043) on the Lab, developed a

pretext/persona (TA0005), and kept trying (TA0003) to secure collaboration. The student's mapping again suggests that pretext (knowing the victim) as well as evasion of scrutiny were central to this objective. This observation can be used by defenders against these types of threats, particularly by understanding what types of OSINT adversaries may seek and how they could acquire it.

Table 6: ATT&CK techniques identified by student teams for Flag 12

Technique ID	Tactic ID	Total	G1	G2	HS
T1036: Masquerading	TA0005: Defense Evasion	2	X	X	X
T1589: Gather Victim Information	TA0043: Reconnaissance	1			X
T1593: Search Open Websites/Domains	TA0043: Reconnaissance	2	X	X	X
T1598: Phishing for Information	TA0043: Reconnaissance	1			X
N/A	TA0003: Persistence	1	X	X	

4. Discussion

4.1 Limitations

Perhaps the biggest limitation is the low number of competing students, which affects the generalizability of these findings. Another noteworthy point is that students might use techniques they are most comfortable with. This is compounded by the temporal constraints of the competition that may limit a more thorough exploration of the tactics and techniques. However, this exploratory effort still offers a good abstraction of how social engineering techniques can be mapped to the tactics and techniques in the framework, which are discussed below.

4.2 ATT&CK tactics, techniques, and social engineering

This paper attempted to map social engineering strategies used by students during a real-time cybersecurity exercise to the ATT&CK framework. Table 7 provides a summary of the tactics and techniques used by students in the social engineering pen testing competition.

Table 7: Overview of tactics and techniques used by students in the competition

1. TA0043: Reconnaissance	2. TA0042: Resource Development	3. TA0001: Initial Access	5. TA0003: Persistence	7. TA0005: Defense Evasion	11. TA0009: Collection	12. TA0040: Impact
Gather Victim Identity Information	Acquire Infrastructure	Phishing	Create Account	Masquerading	Data from Local System	Data Manipulation
Gather Victim Org Information	Establish Accounts	Trusted Relationship			Screen Capture	
Phishing for Information	Stage Capabilities	Valid Accounts				
Search Open Websites/Domains						
Search Victim-Owned Websites						

Table 8 provides a summary of the most popular tactics identified by students. Given the theme of social engineering, it is not surprising to see the extensive use of TA0043: Reconnaissance, TA0042: Resource Development, TA001: Initial Access, and TA0003: Persistence, as they are essentially the preparatory work of OSINT, pretext and persona development, and studying the target.

One tactic, TA0005: Defense Evasion, specifically technique T1036: Masquerading was used extensively by students. This technique is described as occurring "when the name or location of an object, legitimate or malicious, is manipulated or abused for the sake of evading defenses and observation." The description also states that the technique "may include manipulating file metadata, tricking users into misidentifying the file

type, and giving legitimate task or service names." Students may not have considered this latter portion of the description and may have interpreted the first part of the description as applicable to the development of pretexts and personas, which were manipulative and helped evade defenses and observation.

There were a few tactics that appear later in the framework, namely TA0009: Collection and TA0040: Impact. Interestingly, Collection was used for flags 6 (get the dataset) and 10 (change website). In the first case, the specific technique of T1005: Data from Local System was used, which is expected given the objective of getting the dataset from the Lab. In the second case, students used T1113: Screen Capture to take a photo of the page that needed updating. Given that this tactic is often used before 'exfiltration', which applies to obtaining the dataset, but it could also precede the Impact, which was the Lab website posting incorrect (mis) information.

Table 8: Seven most popular tactics used by students

ATT&CK Tactic ID	Total
TA0043: Reconnaissance	66
TA0005: Defense Evasion	35
TA0003: Persistence	19
TA0042: Resource Development	13
TA0001: Initial Access	11
TA0009: Collection	3
TA0040: Impact	1

Table 9 provides details on the most popular techniques identified by students in their playbooks. Currently, the framework has 10 techniques for TA0043: Reconnaissance. Students used 50% of the techniques in this tactic. There are 7 techniques for TA00042: Resource Development and students used 43% of these. Students used 33% of the 9 techniques in TA0001: Initial Access. For the remaining tactics/techniques, students only used 1 or 2 techniques. However, it is interesting to note that the frequency of technique usage did not necessarily align with the number of techniques used. For instance, T1036: Masquerading (TA005: Defense Evasion) was the most used technique even though it was just one of 40 (2.5%) available techniques for this tactic.

Table 9: Nineteen most popular techniques used by students

AT&CK Technique ID	Total
T1036: Masquerading	35
T1593: Search Open Websites/Domains	28
T1598: Phishing for Information	16
T1589: Gather Victim Information	11
T1199: Trusted Relationship	8
T1585: Establish Accounts	6
T1583: Acquire Infrastructure	5
T1591: Gather Victim Org Information	5
T1594: Search Victim Owned Websites	5
T1224: Access leadership areas of interest	4
T1593.001: Social Media	3
T1113: Screen Capture	2
T1566: Phishing	2
T1608: Stage Capabilities	2
T1005: Data from Local System	1
T1078: Valid Accounts	1
T1136: Create Account	1

AT&CK Technique ID	Total
T1565: Data Manipulation	1
T1591.001: Determine Physical Location	1

4.3 A deeper dive into student mappings

The competition allowed for a glimpse into how students were mapping their playbooks to the framework. A few observations can be made:

1. The same flag can be accomplished through similar tactics and techniques despite variations in the pretexts/personas and OSINT gathered. This is expected given the nature of the flags, which required students to do OSINT and develop pretexts and personas. It also highlighted the creativity and human aspects of cybersecurity.
2. Some students may have converged several techniques that may have an overlapping meaning. For instance, when creating personas, students may have used T1036: Masquerading, T1585: Establish Accounts, or T1136: Create Accounts.
3. On several occasions, students used a Tactic but did not identify the specific technique that was used. This was mostly the case for TA0003: Persistence. Furthermore, there were also some occasions when students were clearly using specific tactics and techniques but failed to identify them in their playbooks.
4. Tables 15-17 demonstrate that the majority of the tactics and techniques used were part of the preparation work needed to start the various flags/objectives. Thus, students mostly used social engineering related tactics and techniques, which aligns with the competition theme and emphasizes the human-social-behavioral aspects of cyberattacks.
5. Some tactics and techniques were rarely listed in playbooks. Of particular interest are the tactics and techniques that appear later in the chain. For instance, TA0009: Collection (T1005: Data from Local System and T1113: Screen Capture) and TA0040: Impact (T1565: Data Manipulation) were hardly used. This may be an indication that students did not get far enough in their attempts to pursue the flags and hence did not list them in their playbooks.

4.4 Implications for mitigation

Currently, mitigation techniques identified in the ATT&CK framework do not offer much insight into solutions for social engineering techniques. Given the above analysis, an important point for discussion is what these mappings imply for mitigations against social engineering attacks. Given the detailed use of TA0043: Reconnaissance, TA0042: Resource Development, TA0001: Initial Access, and TA0003: Persistence, employee security training programs might spend more time on these specific tactics. The variety of techniques used in these preliminary tactics in the intrusion chain might further help in designing the training programs. For instance, if T1589: Gather Victim Information and T1591: Gather Victim Org Information are extensively used, training programs may emphasize limiting information that employees and companies make available publicly to limit OSINT findings. Consider for instance this excerpt from G3's report, which aligned their mitigation recommendations with specific techniques:

> The mitigation for Search Open Websites/Domains, Technique (T1593) has no easy mitigation other than control the information that individual can control outside of the organization such as social media privacy settings (e.g., public, private, friends only) and webpage settings (e.g., password protected). The masquerading technique (T1036) by tricking users into misidentifying the file type and giving legitimate task or service names can be mitigated by assigning an authentication code or number to the legitimate task, service, or persona (e.g., social security number). Phishing for Information using a spearphishing link (T1598.003) can be mitigated using anti-spoofing, email authentication, and cyber threat intelligence (CTI) mechanisms along with training to detect malicious links. Establish accounts, techniques (T1585) where adversaries create accounts (e.g., social media) has no easy mitigation other than threat intelligence of these accounts are potentially linked to bad actors.

These suggestions from students could be worked into training programs to make them more effective.

4.5 Value of mapping exercise to students and next steps

One of the biggest advantages the mapping exercise provides is that students can use it as an aid; it allows them to understand where they are in the intrusion chain and how their techniques have helped them move from (or not) one tactic to another to get closer to their objectives. These insights are invaluable to organizations working towards understanding and defending adversary campaigns. Mapping exercises (as well as real threat intelligence) provide insights into the tactics and techniques utilized by adversaries, which can be paired with defensive countermeasures that inform security decisions and operations.

Another important finding is that some of the flags' objectives/impacts may not be represented in the framework as they are intangible or do not align with the philosophy of how ATT&CK techniques are built. Consider for instance, flag 9 (getting hired - insider threat). This flag was accomplished by a few teams and is a legitimate threat for almost all organizations, but the concept of insider threats is not explicitly modeled into the current ATT&CK tactics and techniques. In this case, augmenting ATT&CK or using a use-case specific framework (such as the CTID Insider Threat TTP Knowledge Base) could help capture the nuances highlighted by the various student adversary playbooks. Similarly, flag 2 required the CARE Lab's Twitter account to Retweet a tweet posted by a team (who is posing as someone else). Retweeting a post from a bad actor could damage a company's reputation, which is another intangible harm.

This competition case study offered a first step into understanding how social engineering techniques could be mapped to the ATT&CK framework. For future competitions, the authors could implement the following changes. The competition theme could vary to test out how this influences the mapping process. Students could be given thorough mapping examples during the competition orientation, which would allow them to map more effectively. Students could also receive cheat sheets to make the mapping process easier and more efficient.

The authors hope other educators find this competition case study useful and a template for how to design their own unique and creative course projects that teach students about the wonderful ATT&CK framework and the mapping process.

5. Acknowledgements

This work was supported by the National Science Foundation Award # 2032292. The authors thank their university's ethics board for guiding us to ensure that this workshop was be safe, ethical, and fun for all those involved. The authors also thank their university's Risk Management Unit for preparing all the forms necessary for this workshop.

References

Bleiman, R., Williams, J., Rege, A. & Williams, K. (2022). "Exploring the MITRE ATT&CK® Matrix in SE Education". Proceedings from the IEEE Cyber Science Conference.

Mitrecorp. (2018). MITRE ATT&CKcon 2018: ATT&CK as a Teacher. Available at: https://youtu.be/4s3pZirFCPk).

Nickels, K. et al. (2019). "Getting Started with ATT&CK," ATT&CK. Available at: https://medium.com/mitre-attack/getting-started/home.

Rege, A. & Bleiman, R. (2022). 'Collegiate Social Engineering Capture the Flag Competition', *Proceedings of the 2021 APWG eCrime conference*.

Splunk (2019).. Boss of the SOC (bots) advanced apt hunting companion app: Now available on splunkbase, Splunk. Available at: https://www.splunk.com/en_us/blog/security/boss-of-the-soc-bots-advanced-apt-hunting-companion-app-now-available-on-splunkbase.html (Accessed: February 14, 2023).

Role of Techno-Economic Coalitions in Future Cyberspace Governance: 'Backcasting' as a Method for Strategic Foresight

Mari Ristolainen
Finnish Defence Research Agency, Riihimäki, Finland

mari.ristolainen@mil.fi

Abstract: In an increasingly complex threat landscape, many nations struggle with developing and implementing effective cybersecurity policies for cyberspace governance at a national and international level. Balancing between the demands for establishing national sovereignty and strengthening international collaboration in cyberspace have become a problematic assignment. Collaborating with nations supporting extensively dissimilar ideologies and cybersecurity policies is controversial. Yet, it is almost impossible for a single country to achieve 'self-sufficiency' in cyberspace. Thus, in order to remain competitive, protected, and resilient one must either join or strengthen a developing techno-economic coalition with similar national cybersecurity policies and/or ideological framework. Consequently, this paper argues that techno-economic coalitions serve as an emerging issue or trend for strategic foresight in cyberspace governance in the future. This paper discusses the potential formation of techno-economic coalitions and shows how 'backcasting' can be used in strategic foresight. In this paper, 'backcasting' is not used as a method for creating a traditional strategic map to a future goal, but as a framework for determining what should have happened in order for the techno-economic coalitions to emerge in future cyberspace, i.e. for finding issues or trends that should be followed in strategic foresight today. Firstly, cyberspace governance in relation to national cybersecurity policies is contextualised. Secondly, the concept of techno-economic coalition is defined and the potential emerging techno-economic coalitions are explicated. Thirdly, 'backcasting' as a method for strategic foresight is described. Fourthly, the results of a 'backcasting' experiment in a strategic foresight workshop are presented. And finally, the future formation and role of techno-economic coalitions in cyberspace governance and in cyber defence both at a national and international level are discussed. The role of techno-economic coalitions in future cyberspace governance should be understood and considered today when developing strategic plans and implementing national and international cybersecurity policies.

Keywords: National Cybersecurity Policy, Cyberspace Governance, Cyber Defence, Techno-Economic Coalitions, 'Backcasting', Strategic Foresight, Strategic planning

1. Introduction

Many of the security problems in cyberspace can be considered territorial or national (cf. Hare 2010, 214-215; Kovács 2018, 113-114). Contemporary cyber threats are increasingly targeted at national critical infrastructure, economic competitiveness, national security and citizens, both directly and indirectly. Consequently, many nation-states are interested in the safekeeping of their own national cyberspace and their own national systems. The national controllability of cyberspace and the independent defence of various systems add to the sense of cybersecurity (Rautava & Ristolainen 2022, 239-240). However, the cyberspace governance is often not included in national cybersecurity policies. Cyberspace governance requires international and non-territorial cooperation on both policy and technical levels. Therefore, in the increasingly complex threat landscape, many nations struggle with developing and implementing effective cybersecurity policies for cyberspace governance that is leading towards cyberspace territorialisation.

Balancing between the demands for establishing national sovereignty and the strengthening of international collaboration in cyberspace has become a problematic assignment. Moreover, collaborating with nations supporting extensively dissimilar ideologies and cybersecurity policies is controversial. Yet, it is almost impossible for a single country to achieve 'self-sufficiency' in cyberspace. Here 'self-sufficiency' stands for total technical isolation and technical independence over national critical infrastructure, and is to some extent something larger than legal 'sovereignty in cyberspace' (e.g. Liaropoulos 2013, 21-23; Heller 2021, 1498-1499). Thus, in order to remain competitive, protected and resilient one must either join or strengthen developing techno-economic coalitions with similar national cybersecurity policies and/or ideological framework. Consequently, this paper argues that techno-economic coalitions serve as an emerging issue or trend for strategic foresight in the cyberspace governance of the future. This paper discusses the potential formation of techno-economic coalitions and shows how 'backcasting' can be used in strategic foresight. In this paper, 'backcasting' is not used as a method for creating a traditional strategic map to a future goal, but as a method for determining what should have happened in order for the techno-economic coalitions to emerge in future cyberspace, i.e. for finding issues or trends that should be followed in strategic foresight today.

Firstly, cyberspace governance in relation to national cybersecurity policies is contextualised. Secondly, the concept of techno-economic coalition is defined and the potential emerging techno-economic coalitions are explicated. Thirdly, 'backcasting' as a method for strategic foresight is described. Fourthly, the results of a 'backcasting' experiment of a strategic foresight workshop are presented. And finally, the future formation and role of techno-economic coalitions in cyberspace governance and in cyber defence are discussed. This paper claims that the role of techno-economic coalitions in future cyberspace governance should be understood and considered when developing strategic plans and implementing national cybersecurity policies at present.

This paper is exploratory and belongs to the field of strategic foresight that allows more a creative envisioning of potential futures than maybe more traditional academic papers do. Hence, the developments and scenarios presented may to some seem speculative, but it is crucial to develop sustainable methods and frameworks and to preserve the forward-looking mindset, especially within security organisations (cf. Habegger 2022).

2. Cyberspace governance and national cybersecurity policies heading towards 'cyberterritories'

In previous studies (Ristolainen 2021; Rautava & Ristolainen 2022), it has been estimated that cyberspace governance and national cybersecurity policies are progressively heading towards a formation of a so-called 'cyberterritory', i.e. cyberspace is increasingly considered something similar to territorial waters, airspace and/or land area, i.e. the physical areas over which a sovereign state has jurisdiction over. 'Cyberterritory' can be defined as "an entity of networks and technical infrastructure containing services and the data processed in them that are controlled by a sovereign nation-state" (Ristolainen 2021; Rautava & Ristolainen 2022, 243). This development is seen in the change of cyberspace governance models and national cybersecurity policies that aim for particular state 'sovereignty'[1] in cyberspace (ibid.). Essentially, cyberspace governance models are issues of international law that are being resolved within the UN. Nevertheless, the discussions within the UN show that sovereignty and national controllability of cyberspace are emerging within many nations (ibid.). However, sovereignty in legal terms differs, to some extent, from technical 'self-sufficiency' in cyberspace, i.e. a total technical isolation and technical independence over national critical infrastructure.

Therefore, only a few (if any) nation-states could have necessary resources to be truly 'self-sufficient' in cyberspace. Those who could have the resources (e.g. the United States), don't have the need or desire to be 'self-sufficient' in cyberspace. The nation-states that want to be 'self-sufficient' in cyberspace – wish to gain the same position as the United States (or to deny it) – in a political, a technological, a software infrastructure and a standardization level. The aims can be limited to the control of one's own 'cyberterritory' or to the wider control of global information flows. For these countries the closing of 'the national segment of the Internet'[2] (as, for instance, Russia is planning to) and 'self-sufficiency' is reachable and could work in the time of crisis (Kukkola 2020; Kukkola 2021).

When assessing potential future developments, it seems possible that cyberspace governance and national cybersecurity policies are heading increasingly towards 'cyberterritories' that will eventually lead to a formation of new techno-economic coalitions. In order to remain competitive, protected and resilient national 'cyberterritories' must either join or strengthen developing techno-economic coalitions with other nation states sharing similar cybersecurity policies and/or ideological framework.

3. Techno-economic coalitions

'Techno-economic coalition' is an emerging concept (cf. MGIMO 2019; Ristolainen 2021; Rautava & Ristolainen 2022) that can be defined as: "a coalition formed by separate cyberterritories that develop new coalitional level services, which are based on technologies innovated on coalition level, and thus, promote and benefit the

[1] Cyberspace is associated with very different conceptions of sovereignty - often referred as 'digital sovereignty' (Pohle & Thiel, 2020). However, by 'digital sovereignty' is referred frequently to national regulation of data mobility, i.e. 'data sovereignty' (Braud et al. 2021). In one context, 'digital sovereignty' refers to much broader 'information sovereignty' (see, e.g. Yefremov 2017), and in another context, 'digital sovereignty' denotes a 'national segment of the Internet' that can be disconnected from the global network (see, e.g. Kukkola 2020). Thus, there is no single conception of what sovereignty in cyberspace means, although the same concept 'digital sovereignty' is used (more, see, e.g. Rautava & Ristolainen 2022).

[2] The national segment of the Internet can be defined as "a portion of the Internet infrastructure and services which resides on a state's territory and under sovereign jurisdiction" (Kukkola 2020, viii).

economy of the coalition". Techno-economic coalitions could solve the problem of ideological division. Collaborating with nations supporting similar ideologies and cybersecurity policies would be straightforward.

The role of data is central in the differentiation of techno-economic coalitions. Techno-economic coalitions would share certain values and ideology and distinguish themselves from other coalitions by different attitudes towards (personal) data usage and protection. Where other coalitions would, for instance, use data freely for commercial purposes, others for political and/or religious control, there other coalitions would protect (personal) data profoundly. Data would stay within the techno-economic coalition and be protected (confidentiality, integrity, and availability) by the techno-economic coalition. Data security requirements could vary widely between different techno-economic coalitions. The growing role of data is something that should be taken into consideration already in contemporary strategic planning (e.g. choosing cloud services for national sensitive data).

Potential techno-economic coalitions could emerge around the 'Anglosphere' led by the United States; around China; around Russia; around the European Union; around Islamic countries; around criminal and anarchical groups, and (possibly) around a so-called 'wild card' country or commercial consortium (cf. original listing (1-4) MGIMO 2019; additions (5-7) author). In the following, the most reasonable techno-economic coalitions are categorised[3] in more detail.

3.1 America first and others will follow

The 'Anglosphere' techno-economic coalition would be led by the United States. It would be formed around the USA, Canada, United Kingdom, Australia and New Zealand that are economically tightly integrated and their alliance could also be attractive to countries such as Mexico (MGIMO 2019). This coalition would use its privileged position and geographical location in the world to create the best conditions for itself. This techno-economic coalition would use data freely for commercial purposes and the person would be a product of the business both legally and technically.

3.2 Digital silk road

Similarly, China would be expanding its coalition with neighbouring countries and the Middle East and Africa – tying them to China's economy and infrastructure. The Chinese model would be based on absolute self-sufficiency and it would have access to enormous markets that are largely closed to their competitors' technology and user data. Digital silk road techno-economic coalition could, for instance, export services for network surveillance and social scoring. Data would be controlled by the state and used for commercial purposes.

3.3 Digital Soviet Union

Obviously, Russia wishes to remain an independent global actor and this would be possible only as part of some kind of techno-economic coalition. Russia's coalition would be dependent on the domestic market and public investments. Russia's rational partners would be the members of the Eurasian Union, i.e. the individual states of the former Soviet Union and members of the Collective Security Treaty Organisation (Belarus, Azerbaijan, Kazakhstan, Uzbekistan, Tajikistan and Moldova). The Digital Soviet Union techno-economic coalition would rely on the nostalgia of past ideology and empire-building and use state owned data for ideological control.

3.4 European data, technology and infrastructure idealists

The EU's Digital Strategy of 2020 states that the EU must strengthen its digital sovereignty and set standards instead of lagging behind others (Shaping Europe's Digital Future 2020). EU's idealistic coalition would focus on data protection, technology and infrastructure. The aim would be to strengthen Europe's technological capacity, independence and confidence, and to improve Europe's position in global competition. Data would be owned by the person and data protection and privacy requirements would be extremely high. Still, the future of the EU's techno-economic coalition seems rather uncertain – at the most – idealistic.

[3] This categorisation and the descriptions were enhanced in the research workshop organized by the Finnish Defence Research Agency (FDRA) in 2022.

3.5 Halal Internet

Islamic countries could form a so-called 'Halal Internet' based on shared religious values. This techno-economic coalition would protect the Islamic values from the influence of 'Western culture' and provide digital 'halal', i.e. 'permissible' services, that is, for instance, localized email services and search engines that promote and protect their religious values. Data would be owed by the state and used for controlling what is permitted and forbidden. The success and appeal of this techno-economic coalition could be based on selling religious content filtering and 'permissible services' also for other religions than Islam.

3.6 Where there is light, there must be shadow

The 'shadow' techno-economic coalition would be loosely formed by criminal and anarchic groups. Criminal and anarchic groups could join their forces into a single techno-economic coalition that could try to operate in-between the other techno-economic coalitions and to maintain a 'black-market' for services. The success of this techno-economic coalition would be based on operating outside of any rules and/or laws.

3.7 Wild card

The formation of the 'wild card' techno-economic coalition would happen unexpectedly, perhaps even shockingly. The 'wild card' techno-economic coalition formed by a country or commercial consortium could be a coalition formed around a country of 'sudden success'. The success could be based on, for instance, a discovery of a new raw material (commodity) or a disruptive technological breakthrough.

In the distant future, all the aforementioned techno-economic coalitions could be based on competing national or coalitional technical standards and solutions. Furthermore, techno-economic coalitions could develop new technologies at a national or coalitional level and build services based on national and/or coalitional technology. Nevertheless, the factual occurrence of techno-economic coalitions remains to be seen. Strategic foresight could provide frameworks to analyse the processes that could cause the techno-economic coalitions to emerge factually.

4. 'Backcasting' and strategic foresight

Strategic foresight aims to explore different plausible futures that could arise, and to foresee the opportunities and challenges the future developments could present (Habegger 2022; Kuosa 2012). In this paper, the aim is to find a systematic and structured framework for the estimation of the emergence of techno-economic coalitions. 'Backcasting' is similar to reverse-engineering that works backwards to reconstruct a selected machine or software. Simply, 'backcasting' is a process starting from a 'proposed future' and looking back to contemporary moment to identify the most strategic steps or actions necessary for achieving that specified future (Quist 2016, 125).

In this paper, 'backcasting' is not used as a method for creating a traditional strategic map to a future goal, but as a method for determining what should have happened in order for the techno-economic coalitions to emerge in future cyberspace. In this case, 'backcasting' starts by defining an event that will occur in the future (the emergence of techno-economic coalitions) and then asking the question 'how did this event come to be…?'. The task is then to develop a scenario (or series of events) that explain how the 'proposed future' might actually come about. And finally, the results can be used for finding issues or trends that should be followed in strategic foresight today.

A 'backcasting' experiment offers a way to get a group to envision a 'proposed future' and then to determine together what must happen in order for the 'proposed future' to become reality. Nevertheless, it needs to be highlighted that the results of one particular 'backcasting' experiment serve only as an example of the issues or trends that should be followed in strategic foresight today. Every 'backcasting' experiment produces different results that depend on the people taking part in the experiment. Therefore, 'backcasting' events should be regularly repeated with a different set of experts in order to increase validity.

5. 'Backcasting' experiment as an example

In the spring of 2022, a 'backcasting' experiment for military and civilian experts was held at the Finnish Defence Research Agency (FDRA), where the aim was to test 'backcasting' for strategic foresight. The main task of the FDRA 'backcasting' experiment was to determine which events and changes lead to the appearance of techno-

economic coalitions, i.e. to explain how the 'proposed future' might actually come about. The 'backcasting' experiment started by defining the 'proposed future'. Secondly, the key factors and pathways (keywords) were defined. Thirdly, the identified key factors and pathways were developed into sentences and series of events. Finally, the sentences were summarised into a complicated scenario that was tested from a strategic foresight point of view, i.e. the scenario was used for finding issues or trends that should be followed in strategic foresight today.

The 'proposed future' was defined as a situation where technically delimited 'cyberterritories' have occurred in cyberspace. 'Cyberterritory' refers to an area in cyberspace over which a sovereign state has jurisdiction, i.e. 'cyberterritories' have similar power structures to nation states. Global information flows are controlled at a national level. Many countries are pursuing self-sufficiency and digital sovereignty, but in order to remain competitive they need to either join and/or strengthen the techno-economic coalition. Techno-economic coalitions begin to emerge and they develop new technologies at a national and coalitional level and build services for the coalition based on national/coalitional technologies. Techno-economic coalitions gradually begin to challenge military alliances.

In order to find the key factors and pathways, an adapted model of the PESTEL analysis was used as a supporting research tool. PESTEL is a framework for analysing trends, developments and phenomena in the political, economic, social, technological, environmental and legal environments (cf. Marmol 2015). The experiment applied a PESTEL-M framework to assess political, economic, social, technological, environmental, legal and military factors that could have led to the proposed future. The participants of the 'backcasting' experiment were asked to describe cumulative factors and pathways under each PESTEL-M category that they regarded as necessary for the potential realisation of the 'proposed future'.

In the first phase, the following basic key factors and pathways (keywords) were defined by the participants of the experiment (Table1).

Table 1: 'Backcasting' experiment (PESTEL-M): 1st Phase: basic key factors and pathways (keywords)

Political	– a global confrontation continues – federalism – federal development – the right to privacy – the monitoring of people – the regulation of data usage
Economic	– successful digital currency – global tax agreement – decentralized production – platform economy – remote work
Social	– anonymity (you can be whoever) – inequality – isolation
Technology	– encryption technology changes – fusion energy – data processing and storage – metaverse
Environment	– climate refugees – the reduction of living space – diseases
Legal	– the first virtual person legalised – new types of crimes, e.g. 'cyber homicide' – cyberweapons are legally regulated
Military	– fighting cyber war becomes possible – wars without human casualties – asymmetry – cyber deterrence

In the following phase, the identified basic key factors and pathways were reduced and developed into more complex sentences and series of events that could have led to the formation of techno-economic coalitions (Table 2).

Table 2: 'Backcasting' experiment (PESTEL-M): 2nd Phase: sentences and series of events

Political Development	– Europe (and/or Islamic nations) form a federal state. – A scandal related to data-breach etc. leads to the break-up (disintegration) of the UN. – The need to control personal data grows extensively. – Politicians use the political situation to promote different techno-economic coalitions. – Different techno-economic coalitions use the political atmosphere to distinguish themselves in relation to how individuals can use their personal data and to control how the personal data is collected, stored and used.
Economic Development	– Some form of 'big' digital currency will actually succeed and retain its value (e.g. Digital Dollar, Digital Euro etc.). – Tax unification, i.e. a tax treaty that breaks traditional physical borders between nation states.
Social Development	– It is possible to work remotely globally throughout your career. – Digital isolation is increasing and digital hermits, i.e. persons that exist only in the digital world, are ordinary. – The value of work changes significantly.
Technology Development	– Contemporary encryption technology collapses suddenly (e.g. someone develops a material etc. that causes this) and it leads to the scandal that causes the UN to break-up. – Fusion energy is concretely in sight, i.e. the first devices are already being built. – The processing and storage of large amounts of data evolves, i.e. there are agreements about data storage. – Data becomes available to everybody.
Environmental Development	– A natural disaster causes a large country to become almost uninhabitable and the situation is solved within a techno-economic coalition. – The viable area on earth shrinks.
Legal Development	– The legal definition of 'person' is changed. – The birth of the first fully virtual person is legally possible. – A person other than a legal person can commit a crime. – New types of crimes, e.g. 'cyber homicide', emerge. – Cyber warfare (or a cyber operation that caused significant damage) is tried in court.
Military Development	– To fight cyber warfare becomes possible, i.e. cyber operations cause comparable harm to kinetic operations (cf. legal factors). – Critical infrastructure is defended by military forces. – Wars can be fought without human casualties. – Cyberweapons are legally regulated (cf. legal factors). – Cyber asymmetry and deterrence are reachable with military allies.

And finally, the sentences were summarised into a complicated scenario that demonstrates the driving forces for techno-economic coalitions. The developed scenario and trends that should be followed in strategic foresight today are as follows:

"Unexpected scandals and/or environmental catastrophes are rather difficult to foresee but obviously to be followed when they occur. Some kind of global data breach affecting almost everybody, but especially the world leaders and decision-makers, is likely to happen sooner than later. In the worst-case scenario, this kind of data-breach would be related to the UN or another global actor. The focus in strategic foresight could be on geopolitical challenges and political transitions of different geographic areas that are potential in forming a techno-economic coalition, i.e. according to the results of this backcasting experiment, it is productive to follow the development of countries heading toward a more federally structured government system.

Successful digital currency could be a result of (or benefit from) federally structured government system, i.e. it is important to follow research or prototypes or deployments of digital national currencies. Development of a successful digital currency could also advance the global tax agreement. Global taxing would make it possible to work remotely globally throughout your career, i.e. the different tax proposals and global tax policies should be followed. Nevertheless, the described development could have rather negative side-effects, such as, digital isolation and digital hermits, i.e. the social factors to be followed could be how the meaning and value of work are changing and how the major transformations in the global labour market are fulfilled.

Legislation and law-making processes always play catch-up with technological development. Therefore, legal issues are often outdated for strategic foresight. Nevertheless, the precedents related to the legal definition of a person, new types of cyber crimes and cyber warfare are important to follow. For instance, discussions of the relationship between natural and artificial persons are ongoing and how the law recognises a virtual person needs to be solved in the future.

Legislation and technological development are deeply related to the utilisation of military capacities. When it is possible to fight a cyber war without human casualties, the targets are in the operating systems of the critical

infrastructure. Therefore, cyber defence in the future could be based on the different values and technical standards and solutions developed nationally and/or at the techno-economic coalition level. Different solutions and technologies would create cyber asymmetry and deterrence between techno-economic coalitions."

This modest 'backcasting' experiment shows that the listed keywords, paths, trends, developments and phenomena may seem at first disconnected and trivial. However, together they form drivers of change that may eventually affect the development of techno-economic coalitions in considerable ways. This experiment demonstrates that 'backcasting' offers a usable framework for strategic foresight.

6. Discussion: Role of techno-economic coalitions in cyberspace governance and in cyber defence

The change in cyberspace governance models reflects a change in how nation-states consider cyber threats more in the national framework than at the global level. This development is affected by the fact that cyber threats are targeted, both directly and indirectly, at national critical infrastructure, economic competitiveness, national security and citizens. Nowadays, many nation-states are interested in the safekeeping of their own national cyberspace and their own national systems. The national controllability of cyberspace and the independent defence of various systems add to the sense of cybersecurity.

Increasingly, the sovereignty and 'self-sufficiency' in cyberspace are seen as an answer to security threats. The desire to use similar international laws, which direct the relations between states in the physical geographical environment, in cyberspace is growing. Eventually, this leads to the formation of 'cyberterritories' that could restrict (or control) everything from data movement and storage to political and/or religious control. Nevertheless, it is almost impossible for a single country to achieve 'self-sufficiency' in cyberspace. In order to remain competitive and protected, one must either join or strengthen emerging techno-economic coalitions.

This paper argued that the role of techno-economic coalitions in the future cyberspace governance should be understood and considered when developing strategic plans and implementing national cybersecurity policies at present. This paper examined what it would take for techno-economic coalitions to factually emerge in the future. 'Backcasting' was implemented to explore concrete changes and actions that are necessary for the techno-economic coalitions to develop. The results of the backcasting were placed into a scenario to find issues or trends that should be followed in strategic foresight today. According to the 'backcasting' experiment the trends are, for instance, the expansion of federally structured government systems, digital national currencies, changes in global and remote labour market, the development of global taxing and different legal precedents related to cyberspace.

Divergent techno-economic coalitions would significantly influence and complicate the international cyber law development and application in practice. Nevertheless, techno-economic coalitions would require a desire to use the classical principles of international law in cyberspace, i.e. all the similar laws that govern state relations in the physical geographical environment. Therefore, techno-economic coalitions are also military and political alliances and have strategic importance. Thus, also cyber defence in the future would be based on the different values and technical standards and solutions of future techno-economic coalitions. Consequently, the role of traditional military alliances in future cyber domain would diminish substantially, which should be considered when doing strategic planning.

Acknowledgement

I wish to acknowledge the expertise of all the participants of the 'backcasting' experiment and colleagues at the Finnish Defence Research Agency (FDRA) whose observant comments enhanced this paper significantly.

References

Braud, A., Fromentoux, G., Radier, B., & Le Grand, O. (2021) "The Road to European Digital Sovereignity with Gaia-X and IDSA", *IEEE Network*, Vol 35, No. 2, pp 4-5.

Habegger, B. (2022) Securing the Future: The Use of Strategic Foresight in the Security Sector. *Strategic Security Analysis*. Geneva Centre for Security Policy (GCSP), January 2022, Issue 23.

Hare, F. (2010) The Cyber Threat to National Security: Why Can't We Agree? *Conference on Cyber Conflict Proceedings 2010*, C. Czosseck and K. Pollins (eds.), CCD COE Publications, 2010, Tallinn, Estonia, 211-225.

Heller, K.J. (2021): In Defence of Pure Sovereignty in Cyberspace. *International Law Studies*, Vol. 97, pp. 1432-1499.

Kovács, L. (2018) National Cybersecurity as the Cornerstone of National Security, *Land Forces Academy Review*, Vol. XXIII, No 2(90), 113-120.

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, Helsinki.

Kukkola, J. (2021) *Rakenteellisen kyberasymmetrian strategiset vaikutukset: Venäjän kansallinen internetsegmentti sotilasstrategisena ilmiönä*, Finnish Defence Research Agency, Riihimäki.

Kuosa, T. (2012) *The Evolution of Strategic Foresight: Navigating Public Policy Making*. Routledge, London & New York.

Liaropoulos, A. (2013) Exercising State Sovereignty in Cyberspace: An International Cyber-Order Under Construction? *Journal of Information Warfare*, Vol 12, No. 2, pp. 19-26.

Marmol, T. (2015) *PESTLE Analysis: Understand and Plan for Your Business Environment*. 50Minutes.

MGIMO (2019) "Mezhdunarodnye ugrozy 2020: Kazhdyi za sebia", *Laboratoriia analiza mezhdunarodnykh protsessov MGIMO MID Rossii*, [online], https://mgimo.ru/upload/iblock/2ac/int-threats-2020.pdf, [Accessed December 19 2022].

Pohle, J., & Thiel, T. (2020) Digital Soverignty, *Internet Policy Review*, Vol 9, No. 4, pp 1-19.

Quist, J. (2016) Backcasting. *Foresight in Organizations: Methods and Tools*, P. van der Duin (ed), Routledge, London & New York, 125-144.

Rautava, J-P., & Ristolainen, M. (2022) Cyberterritory: An Exploration of the Concept, *Proceedings of the 21st European Conference on Cyber Warfare and Security (ECCWS)*, Thaddeus Eze, Nabeel Khan & Cyril Onwubiko (eds.), Chester, Academic Conferences International Limited, 239-246.

Ristolainen, M. (2021) "Softaa kyberrajalle! Katsaus kybertilan valtioalueellistamisprosessiin meillä ja maailmalla", *Tutkimuskatsaus 1/2021*, Puolustusvoimien tutkimuslaitos, [online], https://puolustusvoimat.fi/web/tutkimus/tutkimuslaitoksen-julkaisut#tutkimuskatsaukset, [Accessed January 18 2023].

Shaping Europe's digital future (2020) Luxembourg: Publications Office of the European Union [online], https://ec.europa.eu/info/sites/default/files/communica-tion-shaping-europes-digital-future-feb2020_en_4.pdf, [Accessed January 18 2023].

Yefremov, A. (2017) "Formirovanie kontseptsii informatsionnogo suvereniteta gosudarstva", *Zhurnal Vysshei ekonomiki*, No. 1, pp 201-215.

Digital Streets of Rage: Identifying Rhizomatic Extremist Messages During a Hybrid Media Event using Natural Language Processing

Teija Sederholm, Petri Jääskeläinen, Milla Lonka and Aki-Mauri Huhtinen
Department of Leadership and Military Pedagogy, National Defence University, Helsinki, Finland

teija.sederholm@mil.fi
jaaskelainenpet@gmail.com
milla.lonka@outlook.com
aki.huhtinen@mil.fi

Abstract: This research explores how to identify extreme messages during a hybrid media event happening in a small language area by utilizing natural language processing (NLP), a type of artificial intelligence (AI). A hybrid media event gathers attention all sides of the media environment: mainstream media, social media, instant messaging apps and fringe communities. Hybrid media events call attention for participation and activities both in the physical world and online. On the darker side of media events, the media landscape can act as a channel for all kinds of disinformation, hate speech and conspiracy theories. In addition, fringe communities such as 4chan also spread hate speech and duplicated content during hybrid media events. From theoretical point of view, this connection between the physical world and information networks can be seen as rhizomatic in nature, because information spreads without regard to a traditional hierarchy. The result is that when individuals participate in a big media event, there is a viral awareness of different viewpoints and all kind of topics may be posted online for discussion. In addition, in rhizomatic context different kind of arguments can twist each other, "copy and paste", and create very diversity meanings of new comments. The role of extremist speech in online spaces can have effects in physical world. The focus of this paper is to present the findings of a case study on messages posted online by three different actor groups who participated in demonstrations organized on Finnish Independence Day. In this research, two data sets were collected from Twitter and Telegram and Natural Language Processing (NLP) was used to classify messages using extremist media index labels. Three actor groups were identified as participating in the demonstrations, and they were labelled as: far-right, antifascists and conspiracists. Computational analysis was done by using NLP to categorize the messages based upon the definitions provided by the extremist media index. The analysis shows how AI technology can help identifying messages which include extremist content and approve the use of violence in a small language area. The model of rhizome was valid in making the connections between fringe, extremist content and moderate discussion visible. This article is part of larger project related to extremist networks and criminality in online darknet environments.

Keywords: Extremism, Social Media, Artificial Intelligence, Rhizome

1. Introduction

The media landscape acts as a channel for all kinds of disinformation, hate speech and conspiracy theories. During big media events it is not uncommon for demonstrations to be coordinated by different groups using online resources. Over the last ten years, anarchists, far-right and antifascists have organized demonstrations at events surrounding The Independence Day of Finland. However, at the same time, demonstrations happen also in hybrid media environment including e.g., social media platforms, instant messaging apps and the mainstream media (Sumiala et al, 2018). From a theoretical point of view, this connection between streets and information networks can be seen as a rhizomatic nature: "any point of a rhizome can be connected to any other" or "any point of a rhizome can be fragmented to any other" (Deleuze and Guattari, 1980). The rhizomes show how topics appear and twist the discussion related to the hybrid media event.

The purpose of the research is to study the issues and possibilities of identifying Finnish extremist messaging with natural language processing (NLP) during a large Finnish hybrid media event, the Independence Day. While most of the discussion during a hybrid media event is not extremist, the spreaders of extremist media content can attach to the discussion in online space. The concept of rhizomes (Deleuze and Guattari 1980; 1988) is used to understand this phenomenon. The artificial intelligence is taught with messages labelled with the three-level Extremist media index (Holbrook, 2015). The aim of this research is to show possibilities of using combination of NLP and human analysis in identification of content endorsing violence.

A hybrid media event focuses on the contribution that different actors, messages and platforms make to a media event. These components create hybrid interplay in a complex network between mass media outlets, internet-based platforms and mobile communication technologies. A hybrid media event flows fluently between different actors and platforms creating an event that exists in multiple dimensions at once, connecting through a variety of ways. A hybrid media event consists of five elements; actors, affordances, attention, acceleration and affect. Actors can be both human actors and the platforms themselves. Affordance refers to the way different platforms

can be used for different purposes by different operators. Attention is crucial for getting more exposure for the content. Affects are means to accumulate attention and manage meanings such as strong emotions like fear or hate. Acceleration and movement are critical for hybrid media events, since they cause the content to circulate wider. These five elements interact with each other as a rhizomatic model, and create a basis for understanding how the hybrid media environment works. (Sumiala et al, 2018).

Actors with visibility on social media, e.g. accounts of mainstream media, direct much of the attention during a hybrid media event (Uusitalo and Valaskivi, 2020), though other actors in social media play a significant role as well. In addition, fringe communities, such as the imageboard website *4chan,* can be used to spread hate speech and duplicated content to gather attention (Zelenkauskaite et al, 2020). In addition, fringe communities have a significant role in circulating disinformation and conspiracy theories (Zannettou et al, 2017). Conspiracy theories, disinformation and hate speech can be used by different actors as a part of extremist messaging to dehumanize their 'enemies' and gather attention in hybrid media environment (Stecula and Pickup, 2021).

Hybrid media events can be described with a concept of a *rhizome.* Deleuze and Guattari (1980; 1988) use the term *rhizome* to describe a model of multiple, non-hierarchical entries and exit points in data representation and interpretation. A rhizome is an opposite to an arborescent (hierarchic, tree-like) conception of knowledge working with dualist categories and binary choices. Instead, a rhizome works with flat connections, and at any point of a rhizome it can be connected to any other part of the rhizome. Therefore, a rhizome may be broken at any point, and start up again later in some other lines. Thus, it is not possible to create any structural models, because a rhizome resembles more of a map without tracking. Due to this, actors may not know the end state of their actions and thus messages can have unpredictable but vital effects. Hence, it can be difficult to make differences between threats and opportunities in advance in a rhizomatic environment (Deleuze and Guattari 1988.)

The rhizomatic structure between the extreme and 'normal' content creating and sharing during a hybrid media event could be explained with a metaphor of Lego bricks versus watercolour: In principle, it is possible to extract a brick from a Lego-house without influencing its structure (Röck 2019) whereas in watercolour painting there is no exact or simple way to separate one colour from another afterwards. It is possible to extract the distortions, in this research the extremist media content, from the information flow. However, they stay connected to the normal discussion via rhizomes like colours in a watercolour painting. The rhizomatic nature of a hybrid media event helps those distortions to receive more attention. The actors participate the hybrid media event by commenting, sharing and reacting to events in both physical and digital world. Different actors can use conspiracy theories or mis- and disinformation to spread their own agenda, while attending a discussion about an ongoing topic (Sederholm, Huhtinen and Jääskeläinen, 2020). In this study, the groups participating the demonstrations are resistance movements trying to influence narratives created by officials and mainstream media. The four principles of rhizome Connection, Fragmentation, Escaping presumptions and Main role are used to categorize the extremist media content data.

Extremism could be characterized as an "ideological position embraced by those anti-establishment movements, which understand politics as struggle for supremacy rather than as peaceful competition between parties with different interests seeking popular support for advancing the common good" (Bötticher, 2017). In an extremist worldview, there are only friends and foes. The role of extremist speech in online spaces can have effects in the physical world. For example, online hate speech that include far-right political views can predict offline hate crimes against minorities (Gallacher and Heendric, 2021). This study examines three extremist groups; the far-right, antifascists and conspiracists. The distinctions between terms of right-wing extremism and right-wing radicalism are not clear even between academics, but both include authoritarianism, their anti-democracy, and their exclusionary and/or holistic nationalism with accompanying characteristics of xenophobia, racism, and populism (Carter, 2018). *Conspiracy theories are defined with a conspiracy narrative, that consists of a "group of malevolent people, conspiring to bring about some state of affairs to the detriment of the people in general", and this narrative does not exist outside of the conspiracy theory (Konda, 2019).* The view of antifascists as extremists has been prevailing in recent years around the world. Most of antifascism is seen nonviolent. In the US, the antifascists exercise restraint from violence, but "even if most of their 'everyday anti-fascism' is non-violent, militant anti-fascists use violence in confronting targets" (Copsey and Merill, 2020).

2. Data & Methodology

Two datasets were collected for this research. One from Twitter, consisting of three lists of words identified by researchers beforehand. The lists of words were determined based on the words likelihood to be used in content

related to actors participating in demonstrations. Second dataset was gathered from Telegram, and it consists of ten groups identified as potential places for extremist messaging. These datasets represent two aspects of the hybrid media event: a public social media platform and a closed instant messaging app. The datasets were collected from one day before to one day after the Finnish Independence Day demonstrations. The data was labelled by teaching NLP artificial intelligence to label Finnish messages with extremist media index.

The focus of this paper are the three actor groups participating in demonstrations during the Finnish Independence Day: far-right, antifascists and conspiracists. The level of extremist messages was identified by using the extremist media index (Holbrook, 2015). The index consists of three levels: 1. Moderate, 2. Fringe and 3. Extreme. Level 3 has three sublevels which were excluded from this research. Extreme includes justifications of violence. In fringe, the content must be radical and have hostility towards outgroup, but without justifications of violence. This includes content which directly targets the group by aggressive intention or using names that go beyond traditional political discourse (e.g. racial slurs).

In Finland, about 25% of ages 15-74 uses Twitter weekly (Pönkä, 2021). It is not nearly the most popular platform, but in this paper, it represents a platform with active, easily followed discussions about a hybrid media event taking place. The Twitter data was framed with key words used in Finnish Twitter discussions about the participating groups. The words were categorized as demonstrations, far-right, antifascist, conspiracy theorist and other (Table 1). On Twitter, the data was gathered via Twitter API by searching tweets with a word list including words for the demonstrations and words the actors were likely to use for spreading extremist media content about their opposing sides.

Table 1: Twitter search term examples.

Word group	Description	Search term examples on Twitter
Demonstrations	Referring to different demonstrations and overall words related to demonstrations	demonstration, demonstrator, protest, protestors, march, Helsinki without Nazis, 612, Finland wakes up
Far-right	Referring to far-right actors	far-right, extreme right, nazi, national socialism, national socialist, nationalist, nationalistic
Antifascist/left	Referring to antifascist actors	extreme left, anarchist, woke, leftie, communism, communist, commie, socialism, socialist, antifa, antifascist, antifascism
Conspiracy theorists	Referring to people that believe in conspiracy theories	conspiracy theorist, conspiracy, tin foil hat, denialist
Other	Words that may escalate extremist speech during demonstrations	terrorism, terrorist, terror, terrorist attack, elite, criminal, immigrant, Soviet Union, violence

Among these search terms are some words, which may have a negative connotation that can influence the data. For example, “troll” may refer to real trolls or be a way of downplaying a message by saying it was sent by a troll. The role of the word list “Other” is to bring out potential extreme speech which occurs when mentioning civil movements, terrorism or certain groups of people as a rhetoric tool for spreading extremist ideology. The retweets without quotations were not included to this data, since it would multiply the amount of the actual content. The amount of the final data was 5600 tweets.

On Telegram, the data was gathered from nine groups created for discussion related to far-right, antifascists and conspiracy theorists plus one for overall political discussion. The search was limited to known groups consisting of the users likely to attend the demonstrations and discuss about them, plus a single group for overall political discussion (Table 2). The number of messages in all groups together was 1626.

Table 2: Telegram groups.

Telegram group type	Description of group	Telegram group name (names translated from Finnish)
Far-right	Relation to known far-right movements	Defend Finland, *Backstage*, Soldiers of Odin
Antifascist/left	Relation to known antifascist movements	Helsinki without Nazis, A-group, Antifascist metropolitan area
Conspiracy theorists	Relation to known conspiracy theory activists	Truth seekers discussing, The Future of Finland, Universal Superhero Academy
Other	Group for overall political discussion	Politics

The far-right data was *collected from three different Telegram groups. Defend Finland is a large public discussion channel for activists from multiple groups whereas Backstage is a smaller offshoot group of Defend Finland with*

fewer actors. Finnish Soldiers of Odin's is a chat group. The chosen conspiracist groups were "Truth seekers discussing"*, "Universal Superhero Academy" and "*The Future of Finland"*. They were the most active channels* and were thus identified as places for conspiracy theory discussion.

Finnish antifascists groups found on Telegram were channels used to spread information and not as much to discuss about events. This makes them very different to the other two other categories. We collected the messages from three channels, "Antifascist metropolitan area", "A-group" and "Helsinki without Nazis", since they were likely to spread content related to the demonstrations. There are most likely more Finnish antifascist channels or groups on Telegram, but these are the only public ones we could find.

There were some preconditions used in labelling the data. Content that says something negative without real hostility towards the outgroup, e.g. "the stupid leftists are ruining this country" is excluded from the category fringe. Other messages excluded from fringe are ones where outgroups are used as slurs either towards the individuals of ingroup, since they do not target the actual groups of people. Demonstrations are confrontational by default, which is why these preconditions are used to identify fringe and extremist speech in the context of demonstrations. The distinction between concrete threat of violence and users using violent speech as a metaphor can also be difficult to understand, both for human actors as well as the artificial intelligence. Actors saying "throw Nazis to the ocean" are using a common phrase without a concrete threat of violence. They likely don't mean the messages as a concrete action, but as a locution inside their group. In this paper the solution was to label known catchphrases, such as this, as fringe. In cases where the catchphrase is not commonly used in Finnish activism, this sort of speech was identified as extreme, since it can be qualified as a vague reference to violence. Implicit and vague references to violence were also labelled as extreme.

The computational analysis was done by teaching AI technology created and provided by Aiwo (www.aiwo.ai). One researcher used the extremist media index to label messages collected from both Twitter and Telegram. Text analysis applies novel Natural Language Processing (NLP) technology called Bidirectional Encoder Representations for Transformers (BERT), which is based on deep learning neural networks (Devlin et al, 2019). The methodology introduces common language models that are pre-trained, typically using large corpus of data in context of a selected language. The models are then fine-tuned for specific analysis tasks.

The BERT method represents text as a set of tokens that describes words or pieces of words, their context and order in sentences. Furthermore, the method encodes tokens to vectors which are then fed to a neural network. The training algorithm presents the same data vectors as input and output to the network so that some of the tokens are masked in input, which adapts the neural network to learn the masked tokens in their sentence context. This training process produces a common language model. In analysis tasks the classification layer is attached to the pre-trained language model, and this full neural network is fine-tuned using training data that consists of sample texts and their class labels. (Devlin et al, 2019)

This study used the English pre-trained BERT model (Hugging Face, 2018). For classification, 1698 samples of the tweets were labelled from Twitter to three classes: extreme (75 samples), fringe (347 samples) and moderate (1276 samples). This data was sampled randomly to training and validation sets which were used to fine-tune the classification model and to evaluate the resulted model. The model performance was evaluated using F-score to obtain classification accuracy of the validation data.

Table 3. F-score of validation set with three classes.

Label	True positive	True negative	False positive	False negative	F1
Moderate	129	56	18	9	0.905263
Fringe	42	143	11	16	0.756757
Extreme	5	191	5	11	0.384615

Table 4. F-score of validation set with two classes.

Label	True positive	True negative	False positive	False negative	F1
Moderate	132	27	3	7	0.963504
Fringe/Extreme	27	132	7	3	0.843750

The first approach classified tweets to three classes: moderate, fringe and extreme. Quantitatively, F-score presented in Table 3 shows that the classification has rather large portion of false hits, especially with the extreme class. In the next approach the fringe class was merged to the extreme class which led significantly better result (see Table 4). This implies to two causes. First, the number of extreme class is too small compared other classes, and secondly, the labelling of the fringe class is potentially too ambiguous with the extreme class.

3. Findings

Based on the scores, the model recognizes moderate messages well and is able to differentiate it from other categories, but had issues with separating fringe and extreme content from each other. This issue rose from the amount of data, which was not as large as it should possibly be for teaching the NLP-technology. Even when the data was collected during a hybrid media event with extremists participating in it, they weren't active enough in social media to give sufficient data for modelling.

Based on the computational analysis made by the AI technology, the number of extreme messages was very low and the number of fringe messages was almost as big as the number of moderate messages. The reason for this is that the collection of data was based on controversial and emotion-stirring topic of demonstrations during the Independence Day. What is noticeable is that even if the number of messages was largest during the Independence Day, the relative amount of fringe messages was larger after the event. This may be explained by users reacting to the media event not only in real-time, but after it. The hybrid media environment expands the possibilities of going back in time and space inside the digital environment, which is why the users will react to older topics. The same phenomenon occurred with extreme messages.

The far-right groups had the most active discussion but much of it was due of one provocative actor acting before and after the demonstrations. The use of conspiracist and antifascist chats and groups were different, with both acting mainly as a place to share information in. The actors inside conspiracist chat groups perceived authorities as their outgroup but didn't use extremist messaging to talk about them. The antifascist groups didn't mention their outgroups at all.

The AI was taught to categorize messages as fringe and as extreme from each topic. Through the rhizomatic model, the AI shows what kind of role these messages had during the hybrid media event. The AI identified 11 main topics from the data: Parties and politics, Security, Independence Day celebrations, Society, Citizens, Governance, Traditions and culture, Health security, Relations to outside world, history of Finland and the experience of being a Finn. All of these topics were broken to smaller subtopics. To compensate the fact that the AI technology was not able to distinguish extreme and fringe properly, another two analyses were done with the focus on them. This way the computational method was able to create overall picture of the extreme and fringe messages. Tables 5 and 6 show the top three themes (neo-nazism, extremism, fascism) based on the amount of fringe and extreme messages. They were analysed with their rhizomatic connections.

Table 5. Fringe messages analysed with rhizomatic connections.

Rhizome	Topic: Neo-nazism	Topic: Extremism	Topic: Fascism
Connected	Against Nazism and use of term "Nazi" to describe outgroup regardless of their nazism	Against all extremist groups	Discussions against far-right demonstrators and framing of antifascists as fascists
Fragmented	Users using Nazism to frame their (political)outgroup	Use of "whataboutism" to move discussion away from far-right topics	Break the discussions about far-right with framing antifascists as real fascists
Escape presumptions	Messages calling pro-left and governmental actors Nazis	No mentions of conspiracy theorists	Discussions about antifascism as real fascism
Main role	To oppose the far-right demonstrations	To oppose all demonstrations, though the focus is on framing antifascists as bad as far-right	To oppose far-right and all extreme actors

Table 6. Extreme messages analysed with rhizomatic connections.

Rhizome	Topic: Neo-nazism	Topic: Criminality	Topic: Terrorism
Connected	Against the far-right	Accusations of governmental actors being criminals	Discussions about a plotted far-right terrorist plot and its connections

Rhizome	Topic: Neo-nazism	Topic: Criminality	Topic: Terrorism
Fragmented	Shift the focus of discussions to threat of far-right or the antifascists being 'real Nazis'	Discussions mainly unrelated or vaguely related to demonstrations	To shift the focus into the role of political right in supporting terrorists
Escape presumptions	Some messages using Nazism to describe political left and governmental actions	One individual spamming multiple messages written in same format. This actor also wrote much fringe content.	Some actors accusing antifascists as terrorists based on the situation in USA
Main role	To diminish the human value of far-right and their right to action in society	To create and spread distrust in authorities	To spread accusations of potential extreme actors and their role in supporting terrorism

Based on the analysis, it's possible to teach NLP to categorize Finnish-language messages during a hybrid media event and identify which of those include extreme ideas. The extreme does not automatically mean the messages incite users for violence, but they can be used to justify the use of violence during demonstrations.

4. Discussion

The purpose of this research was to study the issues and possibilities of identifying Finnish extremist messaging with natural language possessing (NLP) during a large hybrid media event, the Finnish Independence Day. The data was gathered from two social media platforms, Twitter and Telegram. Overall, the demonstrations during the Independence Day were mostly peaceful and the number of participants was lower than expected. However, it can still be argued through analysed data that the model of rhizome was valid in making the connections between fringe, extremist content and moderate discussion visible during a hybrid media event.

The results indicate that it is possible to teach NLP to categorize Finnish-language messages during a hybrid media event and to identify extremist text content. It is notable, that extreme content does not automatically mean the messages incite users for violence, but they can be used to justify the use of violence during demonstrations. The identification of the differences between moderate and fringe/extreme messages was very accurate, but the finetuning between fringe and extreme messages was not as successful. As Holbrook (2015) writes, "There are grey areas between categories and some boundaries are difficult to define, particularly between vague fringe material and vague level 1 extreme material therefore a degree of overlap between categories should be assumed". This is most probably due to Finnish language and a low number of extreme messages. It is possible, that a relatively larger data could make the recognition model more accurate.

There can be multiple reasons for the low number of extreme messages, even when part of the data was collected from Telegram groups associated with potential extremist movements. Firstly, the chat groups were publicly known and this may have influenced the nature of discussion. Secondly, the participants were on the streets during the event or celebrating Independence Day elsewhere and thus not so active on social media. Lastly, there was no outgroup present to target with extreme messages. On Twitter, the conversation was more heated with the outgroups being present on the same platform. Also, the role of antifascist Telegram-channels was to send information, not embrace a discussion. There are still some issues to be solved. How to recognize the difference between hostility and pungent critique towards an outgroup? The outgroups such as Nazis and communists are used both to discuss the groups and to stigmatize the actors. It is critical to teach the AI to recognize fringe and extreme messages. Another issue are words with negative connotations, like being able to distinguish the confrontative attitude towards communists from messages of what communists have done historically.

The model of rhizome was valid in making the connections between fringe, extremist content and moderate discussion visible during a hybrid media event. The use of the rhizome as model of data analyse reflects its expansive and ephemeral nature, rather than one which forecloses its conceptual possibilities, is promoted (Lawley, 2005). The extreme flows and blends into non-extreme discussion via rhizomes and gets mixed up in normal political discussions. The concept of rhizome can help to see the connections between extreme and moderate discussions, possibly to even predict where extreme and fringe messages may appear.

The messages analysed were collected afterwards meaning Twitter or Telegram did not label them breaking the terms of the platform. This might mean their ability to identify extremist messages in small language areas such as Finnish is inefficient. In addition, this means that these platforms still allow extremist content to spread even though it breaks their terms of use. If the social media cannot distract the colours of extreme from the watercolour painting of political discussion, it continues to influence radical actors and stir up emotions

regarding controversial topics, such as demonstrations and activism. For future research, the Holbrook's extremist media index can be used as a basis for the learning process for the natural language processing of extremist messaging in small language area.

Acknowledgements

This study is part of a research project funded by the Academy of Finland

References

Bötticher, A. (2017) Towards Academic Consensus Definitions of Radicalism and Extremism. *Perspectives on Terrorism*, *11*(4), 73–77. Available at: http://www.jstor.org/stable/26297896. (Accessed January 19, 2023).

Carter, E. (2018) Right-wing extremism/radicalism: reconstructing the concept, *Journal of Political Ideologies*, *23*(2), 157-182.

Deleuze, G. and Guattari, F. (1988) *A Thousand Plateaus Capitalism and Schizophrenia*, Bloomsbury Academic.

Deleuze, G. and Guattari, F. (1980) A Thousand Plateaus, Trans. Brian Massumi. London and New York: Continuum, 2004. Vol. 2 of Capitalism and Schizophrenia. 2 vols. 1972-1980. Trans. of *Mille Plateaux*. Paris: Les Editions de Minuit.

Devlin, J., Chang M-W., Lee K. and Toutanova K. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, NAACL-HLT 2019: Minneapolis, MN, USA*, Volume 1: 4171-4186

Gallacher, J. and Heerdink, M. (2021) Mutual radicalisation of opposing extremist groups via the Internet, *Preprint* doi:10.31234/osf.io/dtfc5

Holbrook, D. (2015) "Designing and Applying an 'Extremism Media Index'", *Perspectives on Terrorism*, 9(5), 57–68. Available at: http://www.jstor.org/stable/26297434 (Accessed January 19, 2023).

Hugging Face (2018) *BERT base model (uncased)*. Available at: https://huggingface.co/bert-base-uncased (Accessed May 15, 2022).

Jalonen, H., Paavola, J., Helo, T., 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,* 15. 100-111.

Konda, T. M. (2019) *Conspiracies of Conspiracies: How delusions have overrun America,* The University of Chicago Press, Chicago, ISBN-13: 978-0-226-58576-5. doi:10.7208/chicago/9780226585932.001.0001

Lawley, S. (2005) Deleuze's Rhizome and the Study of Organization: Conceptual Movement and an Open Future*, Journal of Critical Postmodern Organization Science,* Vol 3 (4), 36-49. Available at: https://www.academia.edu/download/35484409/Tamara.pdf (Accessed January 19, 2023).

Leppänen, M. (2019) Poliisi: 612-kulkueessa oli 1 500 osallistujaa, äärioikeiston vastustajien mielenosoituksessa 2 700, kiinni otettiin 13 ihmistä, *Yleisradio* 6.12.2019. https://yle.fi/uutiset/3-11106587 (Accessed at November 29, 2021).

Pönkä, H. (2021) Sosiaalisen median katsaus 04/2021. Available at: https://speakerdeck.com/hponka/2021 (Accessed January 19, 2023).

Roozenbeek, J. and van der Linden, S. (2021) Inoculation Theory and Misinformation. *Published by the NATO Strategic Communications Centre of Excellence*, ISBN: 978-9934-564-49-9. Available at https://stratcomcoe.org/pdfjs/?file=/publications/download/Inoculation-theory-and-Misinformation-FINAL-digital-ISBN-ebbe8.pdf?zoom=page-fit (Accessed January 19, 2023).

Sederholm T., Jääskeläinen P. and Huhtinen A-M. (2021) Coronavirus as a Rhizome – The Pandemic of Disinformation, *International Journal of Cyber Warfare and Terrorism.* 11(2): 43–55. doi: 10.4018/IJCWT.2021040104

Stecula D. A. and Pickup M. (2021) Social Media, Cognitive Reflection, and Conspiracy Beliefs, *Front. Polit. Sci.* 3:647957. doi:10.3389/fpos.2021.647957

Starbird, K., Maddock, J., Orand, M.; Achterman, P. and Mason, R. M. (2014) Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing*, In iConference 2014 Proceedings* (p. 654 - 662). doi:10.9776/14308

Uusitalo, N., and Valaskivi, K. (2020) The Attention Apparatus: Conditions and Affordances of News Reporting in Hybrid Media Events of Terrorist Violence, *Journalism Practice*. doi: 10.1080/17512786.2020.1854052

Zannettou, S., Caulfield, T., & De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., Stringhini, G. and Blackburn, J. (2017) The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources, *Proceedings of the 2017 internet measurement conference*, 405-417. doi:10.1145/3131365.3131390

Zelenkauskaite, A., Toivanen, P., Huhtamäki, J. and Valaskivi, K. (2020) Shades of hatred online: 4chan duplicate circulation surge during hybrid media events, *First Monday*, *26*(1). doi: 10.5210/fm.v26i1.11075

An Analysis of Critical Cybersecurity Controls for Industrial Control Systems

Nkata Sekonya and Siphesihle Sithungu
University of Johannesburg, Johannesburg, South Africa

nkatasekonya61@gmail.com
siphesihles@uj.ac.za

Abstract: Industrial Control Systems (ICS) comprise software, hardware, network systems, and people that manage and operate industrial processes. Supervisory Control and Data Acquisition Systems (SCADA) and Distributed Control Systems (DCS) are two of the most prevalent ICS. An ICS facilitates the effective and efficient management and operation of industrial sectors, including critical infrastructure sectors like utilities, manufacturing, and water treatment facilities. An ICS collects and integrates data from various field controllers deployed in industrial contexts, enabling operators to make data-driven decisions in managing industrial operations. Historically, ICS were isolated from the internet, functioning as part of air-gapped networks. However, the efficiency improvements brought about by the emergence of Information Technology necessitated a shift towards a more connected industrial environment. The convergence of Information and Operational Technology (IT/OT) has made ICS vulnerable to cyberattacks. Due to the crucial nature of the infrastructure that ICS manage, cyberattacks against ICS may cause critical infrastructure sectors to experience downtime. This may have a crippling impact on a country's well-being and essential economic activities. Given the proliferation of cyber warfare, cyberattacks against ICS are increasingly significant at present, as was the case during the 2015 attack on Ukraine's power infrastructure, which was successful in causing a blackout that affected over 200 000 persons. The threat actors used malicious software known as "BlackEnergy3", which was created to interfere with the regular operation of the ICS in charge of controlling electrical substations. This was the first known instance of malicious software causing blackouts. In response to increasing cyberattacks against ICS, the SANS Institute, in a whitepaper titled "The Five ICS Cybersecurity Critical Controls", present five critical controls for an ICS cybersecurity strategy. This paper discusses ICS and the increased convergence of IT and OT. The paper also outlines significant cyberattacks directed at ICS. The paper then follows an exploratory research methodology done in response to the Five ICS Cybersecurity Critical Controls to determine the state of ICS literature that can help ICS operators secure their environments in accordance with the framework. Additionally, the ICS Cybersecurity Critical Controls are mapped to the NERC CIP standards, which provide guidance on the security of the Bulk Electric System (BES) and associated critical assets in North America.

Keywords: industrial control systems, critical infrastructure, cybersecurity

1. Introduction

Industrial Control Systems (ICS) play an important role in managing complex industrial processes, including those in critical infrastructure sectors like utilities, manufacturing, and water treatment facilities. ICS comprise networked software and hardware systems that facilitate the management of complex industrial processes by enabling process automation, monitoring, and safety in industrial environments (Asghar and Zeadally, 2019). In the past, ICS were separated from traditional IT networks (Thomas and Chothia, 2020). However, the increased adoption of technologies brought about by the emergence of Information Technology has seen a shift towards a more connected ICS environment. As a result of the convergence of Information and Operational Technology (IT/OT), new cybersecurity issues have emerged (Miller et al., 2021). Remote access to ICS, for example, enables remote management of industrial processes and remote maintenance of industrial equipment. This, however, leaves the ICS environment vulnerable to cyberattacks. The successful breach of ICS can result in the interruption of industrial processes, physical equipment damage, human casualties, and devastating physical repercussions (Alladi, Chamola and Zeadally, 2020). This is demonstrated – for example – by the many attacks on Ukraine's electricity infrastructure, which have resulted in outages (Hemsley and Fisher, 2018). In response to attacks on ICS environments, the SANS Institute published a whitepaper presenting cybersecurity controls most critical for ICS environments. The controls are 1) an implementation of ICS-specific incident response plan(s), 2) the use of a defensible architecture, 3) ICS network visibility and monitoring, 4) secure remote access, and 5) risk-based vulnerability management. The objective of the proposed controls is to guide industrial organisations in creating or enhancing their ICS security program, better protecting ICS environments from cyberattacks. The main objectives and contributions of this paper are as follows: To recommend security controls that conform to the SANS Five ICS Cybersecurity Critical Controls after having conducted an exploratory examination of the present state-of-the-art literature in ICS security. To map the SANS Five ICS Cybersecurity Critical Controls to the NERC CIP standards used in the protection of the Bulk Electric System (BES) and associated infrastructure.

To accomplish the stated objectives, the paper is structured as follows: Section 2 discusses the ICS architecture and the most common types of ICS. Section 3 discusses the convergence of IT and OT and the cybersecurity concerns brought by this convergence. Section 4 discusses major cyberattacks affecting ICS infrastructure. The contribution of this work is Section 5, which examines and aligns the SANS Five ICS Cybersecurity Critical Controls with existing ICS security literature. Section 6 provides a mapping of the ICS Cybersecurity Critical Controls to the NERC CIP standards. Section 7 concludes the paper.

2. Industrial Control Systems

Industrial Control Systems (ICS) comprise networked software and hardware systems that allow for the control and supervision of complex industrial processes (Drias, Serhrouchni and Vogel, 2015). Two of the most common ICS are Distributed Control Systems (DCS) and Supervisory Control and Data Acquisition (SCADA) systems. A DCS enables distributed control within a defined geographic area, whereas SCADA systems are used for centralised control and monitoring of geographically scattered assets (Yadav and Paul, 2021). Building automation systems (BAS), physical access control systems (PACS), and safety systems are additional ICS (Stouffer et al., 2022). Common ICS and their components are covered in this section, including the sectors in which they are typically used.

2.1 SCADA Systems

SCADA systems centralise data acquisition and control of geographically scattered assets and are utilised in distribution systems such as water distribution, oil and gas pipelines, transportation systems, and electrical utility transmission (Stouffer et al., 2022). SCADA systems are made up of software, hardware, and network components that can be classified as 1) field devices, 2) communication systems, and 3) supervisory systems. Field devices comprise sensors and actuators for data collection and physical control, as well as Programmable Logic Controllers (PLCs) or Remote Terminal Units (RTUs) responsible for control logic and communication with the control server (Pliatsios et al., 2020). The use of radio, satellite, or power line-based communication mediums, among others, by communication systems allows for communication between field equipment and supervisory systems (Stouffer et al., 2022). Also included in the communication systems are ICS-specific communication protocols like Modbus and DNP3 (Asghar, Hu and Zeadally, 2019). Supervisory systems include the Master Terminal Unit (MTU), which transmits control data to field equipment and collects and stores data in the Data Historian (Drias, Serhrouchni and Vogel, 2015). Operator interaction with data from the MTU is made possible through the Human Machine Interface (Asghar, Hu and Zeadally, 2019). The design of a SCADA system is shown in Figure 1.

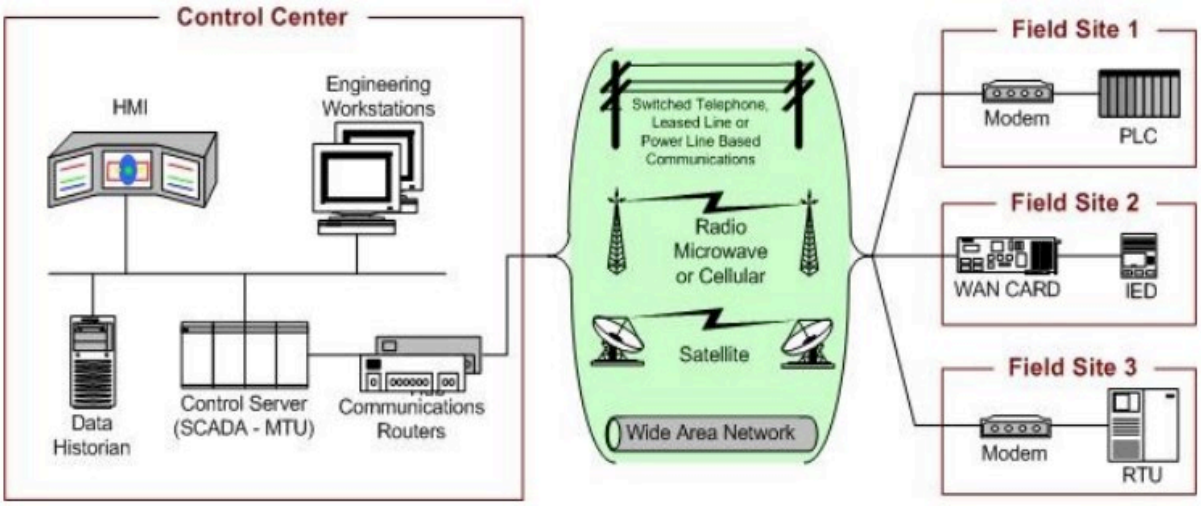


Figure 1: SCADA System General Layout (Stouffer, Falco and Scarfone, 2011)

2.2 Distributed Control Systems

Distributed Control Systems (DCS) are used in the management and control of industrial processes within the same geographic area (Asghar, Hu and Zeadally, 2019). A DCS has a supervisory level of control overseeing multiple controllers distributed across the industrial environment (Bolton, 2021). DCS are typically used in power generation plants, oil refineries and manufacturing (Stouffer et al., 2022). DCS employ most of the components discussed under SCADA systems, with PLCs acting as local controllers. Both SCADA and DCS fall under the

umbrella of Operational Technology. The next section discusses the convergence of Information and Operational Technology.

3. IT and OT Convergence

The Industrial Control Systems used to manage and operate industrial processes fall under Operational Technology (OT). Historically, ICS operated in an environment isolated from enterprise Information Technology (IT) networks and the public internet (Murray, Johnstone, and Valli, 2017). This provided a level of security because, to carry out a successful attack, an adversary had to have physical access to the ICS. While a deterrent, it should be noted that a motivated adversary, such as the threat actors behind Stuxnet, which targeted Iran's Natanz uranium enrichment facility, could still infiltrate such an environment isolated from the internet through infected USB sticks (Miller et al., 2021). Nonetheless, ICS, which traditionally operated in a relatively secure environment, are now connected to enterprise IT and the internet at large. This presents challenges as ICS are not equipped to deal with cybersecurity challenges associated with IT systems. It is in this context that this section discusses the convergence of IT and OT and the resultant cybersecurity challenges.

Operational Technology (OT) systems include SCADA, DCS and associated industrial communication networks. OT is primarily concerned with the safety and consistency of industrial processes, whereas Information Technology (IT) is concerned more with the management of data (Hahn, 2016). The ability to tolerate delays is the first of several distinctions between OT and IT environments that are described in (Stouffer et al., 2022). Because OT powers critical infrastructure sectors, it is less tolerant of delays than IT. In an industrial facility, a delay in the reaction of safety systems might have devastating implications. OT has stricter availability requirements than IT since downtime can have a cascading effect due to the interdependence of critical infrastructure driven by OT systems. Downtime in telecommunications infrastructure, for example, has a negative impact on businesses because the internet is at the heart of many corporate activities.

Another key distinction is that OT prioritises safety above security; as a result, security measures that may jeopardise the safety of OT systems are not put in place. OT operators place a higher priority on the reliability of the industrial process, whereas IT suggests that security controls and upgrades be implemented as soon as they are available. This creates difficulties since OT systems' vulnerabilities may not be fixed for a long time. Systems in OT contexts have fewer resources than their IT counterparts, making it challenging to implement security controls like encryption. The issue is made worse by the usage of proprietary protocols in OT systems, which differ from those used in IT. Applying security rules that are ubiquitous in IT systems becomes challenging or nearly impossible as a result. According to Stouffer et al. (2022), another problem impeding OT system security is reliance on a single OT provider. Third-party security services can invalidate vendor agreements, leaving enterprises operating OT systems with few alternatives for safeguarding their systems. This is especially problematic given that the lifespan of OT components might range from 10 to 15 years (Hahn, 2016).

Cyberattacks on OT environments have escalated, which is a major problem caused by the convergence of IT and OT systems. OT settings are now susceptible to attacks that are widespread in IT systems. This is particularly troubling because OT runs critical infrastructure, whose interruption might have a severe impact on a nation's economy, hinder its capacity to run or even put people's lives in peril. A subset of major cyberattacks targeted at OT systems is discussed in the following section.

4. Industrial Cybersecurity Events

Cyberattacks on Industrial Control Systems may interrupt the operation of critical infrastructure. This is demonstrated by the incidents mentioned in this section. Cyberattacks mentioned in this section are the most significant in terms of their ability to have physical repercussions, disrupt industrial activities, and cause human casualties. This section covers a selection of major ICS attacks.

4.1 Stuxnet

At the time, Stuxnet was the most sophisticated attack on critical infrastructure, and it was used to thwart Iran's nuclear program. In 2010, researchers discovered the Stuxnet malware, which was aimed at Iran's Natanz uranium enrichment complex. Because these facilities were not connected to the public internet, initial access is believed to have been gained through an infected USB stick (Miller et al., 2021). After obtaining access to the network, Stuxnet exploited vulnerabilities in conventional IT systems to infect the PLCs in charge of managing the facility's centrifuges. This allowed Stuxnet to alter the centrifuge speeds there, leading them to malfunction (Firoozjaei et al., 2022).

4.2 BlackEnergy3

In 2015, Ukraine's electricity infrastructure was attacked, resulting in power disruptions for about 200 000 persons (Alladi, Chamola and Zeadally, 2020). BackEnery3 malware was employed in the attack, which was suspected to have been sent using spear-phishing emails (Firoozjaei et al., 2022). The assault used the third version of the BlackEnergy malware, originally detected in 2007, and the KillDisk trojan to wipe out any evidence on targeted systems (Miller et al., 2021). This was the first recorded case of malicious software causing blackouts, and it served as a warning of how devastating an attack on the ICS running critical infrastructure can be.

4.3 Industroyer

In 2016, Industroyer malware, also known as the Crash Override malware, was utilized in another attack on Ukraine's power infrastructure. The initial access is said to have been obtained using spear phishing emails, which are proving effective (Hemsley and Fisher, 2018). This attack was successful in knocking off electricity to 20% of Ukraine's capital, Kyiv (Alladi, Chamola and Zeadally, 2020).

4.4 Triton

Triton is an attack framework that was used in an attack on an ICS at a Saudi Arabian petrochemical company. Triton was used to target a Safety Instrumented System (SIS) in charge of monitoring SIS safety controllers, which are the plant's final line of defence against disasters (Miller et al., 2021). SIS controllers are in charge of avoiding plant failures, and their disruption might have severe consequences for both personnel safety and the surrounding environment. The attack was successful in interrupting facility operations since industrial systems had to be shut down (Hemsley and Fisher, 2018).

Attacks on ICS can result in catastrophic outcomes, as the incidents covered in this section demonstrate. The next section discusses the SANS Five ICS Cybersecurity Critical Controls and looks at recent research to find security measures that can help ICS operators secure their environments in accordance with the framework.

5. SANS Five ICS Cybersecurity Critical Controls

In response to recent cyberattacks on Industrial Control Systems (ICS) around the world, ICS Cybersecurity experts Robert M. Lee and Time Conway from the SANS Institute introduced Five ICS Cybersecurity Critical Controls in their whitepaper titled “The Five ICS Cybersecurity Critical Controls”. The objective of the proposed controls is to guide industrial organisations in creating or enhancing their ICS security program (Lee and Conway, 2022). The proposed cybersecurity controls most critical to ICS systems are 1) an ICS incident response, 2) a defensible architecture, 3) ICS network visibility monitoring, 4) secure remote access, and 5) risk-based vulnerability management (Lee and Conway, 2022). This section discusses the proposed controls and explores the current ICS cybersecurity literature to identify, discuss and map methods that can guide ICS operators to secure their environments in accordance with the proposed framework.

5.1 ICS Incident Response

An incident response plan details the steps taken to respond to a cyber incident, which includes (but is not limited to) preparation, detection, containment, eradication, recovery, and post-incident activity (Staves et al., 2020). (Lee and Conway, 2022) advise that industrial organisations should have an ICS-specific incident response plan. Organisations are advised to begin by identifying scenarios that pose the most risk to their organisation. This process should be intelligence-driven, studying past successful attacks on organisations in the same sector. Secondly, organisations should identify high-consequence scenarios regardless of whether an attack has affected their sector in the past (e.g., ransomware attacks) and determine the steps an attacker would need to take to successfully compromise the organisation’s infrastructure. Lastly, organisations ought to perform tabletop exercises to test the maturity of their incident response plan.

The work of (Miller et al., 2021) provides a comprehensive list of publicly reported cyberattacks aimed at ICS. The study covers attacks which occurred as early as 1988 and as late as 2020. For each cyberattack, the authors discuss the method used by threat actors to gain initial access to the OT environment, the threat actors responsible for the attack, the sector of the targeted organisation and the impact of the attack. This study can assist ICS operators in identifying scenarios that pose the most risk to their organisation. Organisations can study cyberattacks that have in the past targeted their sector. Figure 2 presents a summary of cyberattacks aimed at ICS.

Attack	Date	Initial Access	Threat Actor	Sector	Impact
PLC Password Change	1988	Workstation Compromise	Insider	Manufacturing	Denial of Control
Ignalina Nuclear Power Plant	1992	Workstation Compromise	Insider	Civil Nuclear	Loss of Productivity and Revenue
Chevron Refinery Emergency Alarm System	1992	Workstation Compromise	Individual	Chemical	Loss of Productivity and Revenue
Salt River Project	1994	Internet Accessible Device	Individual	Energy and Water	Loss of Productivity and Revenue, Disk Wipe
Omega Engineering	1996	Workstation Compromise	Individual	Manufacturing	Disk Wipe
Worcester, MA Airport	1997	Internet Accessible Device	Individual	Transport	Loss of Productivity, Revenue, Availability, and Safety
Gazprom	1999	Unknown	Organised Group + Employee	Chemical and Energy	Loss of Productivity and Revenue
Bradwell Nuclear Power Plant	1999	Workstation Compromise	Insider	Civil Nuclear	Disk Wipe
Maroochy Water System	2000	Wireless Compromise	Insider	Water	Damage to Property
Cal-ISO System	2001	Unknown	Nation State	Energy	None Disclosed
Virus on Manufacturing System	2001	Spearphishing	Nation State	Manufacturing	Loss of Productivity and Revenue
Houston Port	2001	Internet Accessible Device	Individual	Transport	Loss of Productivity and Revenue
Gas Processing Plant	2001	Trusted Relationship	Supplier	Chemical	Loss of Productivity and Revenue
PDVSA	2002	Internet Accessible Device	Organised Group	Chemical	Loss of Productivity and Revenue, Disk Wipe
Flight Planning Computer	2003	Unknown	Individual	Transport	Loss of Productivity and Revenue
CSX Train Signalling System	2003	Spearphishing	Unknown	Transport	Loss of Productivity and Revenue
Contractor Infects SCADA Network	2004	Replication Through Removable Media	Unknown	Food	Loss of Productivity and Revenue
Daimler Chrysler Plants	2005	External Remote Service	Individual	Manufacturing	Loss of Productivity and Revenue
Tehama-Colusa Canal	2007	Workstation Compromise	Individual	Water	Damage to Property
Lodz Tram System Hacked	2008	External Remote Service	Individual	Transport	Loss of Safety
US Power Grid	2009	Internet Accessible Device	Nation State	Energy	None Disclosed
Hospital HVAC	2009	Workstation Compromise	Insider	Health	Loss of Safety
Night Dragon	2009	Exploit Public-Facing Application	Organised Group	Energy	Theft of Operational Data
Sality Virus Infects DVS Servers	2009	Unknown	Unknown	Chemical	Loss of View
Stuxnet	2010	Replication Through Removable Media	Nation State	Civil Nuclear	Damage to Property, Manipulation of View and Control
Shionogi	2011	Workstation Compromise	Individual	Health	Disk Wipe
Niagra AX	2012	Internet Accessible Device	Unknown	Manufacturing	Manipulation of Control
Espionage on Iranian CI	2012	Replication Through Removable Media	Nation State	Chemical	Theft of Operational Data, Unintentional Disk Wipe
Turbine Control System	2012	Replication Through Removable Media	Organised Group	Energy	Loss of Productivity and Revenue, Theft of Operational Data
Rye Brook Dam	2013	Internet Accessible Device	Organised Group	Water and Energy	None Disclosed
European Public Utility Services Attacked	2014	Spearphishing	Organised Group	Various	Denial of Service, Theft of Operational Data
German Steel Mill	2014	Spearphishing	Unknown	Manufacturing	Damage to Property
Ukrainian Energy	2015	Spearphishing	Organised Group	Energy	Loss of Productivity and Revenue
Ukrainian Energy	2016	Spearphishing	Organised Group	Energy	Disk Wipe, Loss of Productivity and Revenue, Loss of Safety
Wolf Creek	2017	Spearphishing	Organised Group	Civil Nuclear	None Disclosed
Cadbury Factory Attack	2017	External Remote Service	Organised Group	Food	Loss of Productivity and Revenue
Triton/Petro Rabigh	2017	Workstation Compromise	Nation State	Chemical	Denial of Control, Loss of Safety
Norsk Hydro	2019	Spearphishing	Unknown	Manufacturing and Energy	Loss of View
Triton/Undisclosed	2019	Workstation Compromise	Nation State	Undisclosed	Denial of Control, Damage to Property, Loss of Safety
Hackers Target Oil Producers	2020	Spearphishing	Unknown	Chemical	Theft of Operational Data
Israeli Water Facilities Attacked	2020	Internet Accessible Device	Organised Group	Water	None Disclosed
Cyber-Attack on Shahid Rajaie Port	2020	Unknown	Nation State	Transport	Loss of Productivity and Revenue
Honda Factories Cyber Attack	2020	Spearphishing	Unknown	Manufacturing	Denial of Control

Figure 2: Summary of ICS cyberattacks (Miller et al., 2021)

The work of (Firoozjaei et al., 2022) takes it a step further by analysing the most significant of these attacks and mapping the adversarial tactics and techniques against the MITRE ATT&CK framework. This is crucial, as it provides a mapping of the actions a threat actor would need to perform to successfully compromise an organisation's infrastructure. This enables the organisation to develop incident response plans specific to their ICS environments. Once scenarios posing the most risk to the organisation are identified, and appropriate incident response plans are in place, organisations should implement an architecture that allows them to detect and respond to the identified scenarios. The next subsection discusses the implementation of a defensible architecture.

5.2 Defensible Architecture

This control is concerned with establishing an architecture that satisfies the organisation's operational demands, is adaptive and robust, and supports the data gathering required for security operators to protect against the scenarios outlined in the first control (Lee and Conway, 2022).

Given the growing convergence of OT and IT, a common theme among network architectures is to provide segmentation between the IT and OT network. This is to ensure that attacks on IT do not propagate to the OT environment, affecting ICS. A unified architecture for ICS protection, proposed by (Zhou et al., 2020), offers integrated protection for networks, control systems, and physical processes. The design provides a secure network, secure ICS protocols, and secure end-to-end communications to guarantee the protection of sensitive process data. Securing control systems is done using a risk-based strategy that includes defences based on tolerance and prevention. Additionally, the design offers a defence against process-aware attacks. The next step after establishing a defendable architecture is to guarantee network visibility, which is covered in the next subsection.

5.3 ICS Network Visibility and Monitoring

One of the primary distinctions between Information and Operational Technology (IT/OT) is that IT prioritises data while OT prioritises processes. It is within this context that the visibility of the process in OT environments becomes crucial. Visibility in OT extends beyond individual systems to include the monitoring of ICS-specific protocols (Lee and Conway, 2022). This level of visibility is made possible by the implementation of a defensible architecture, as proposed in the second control, and allows for the collection of data that supports the detection of scenarios identified in incident response planning (Lee and Conway, 2022).

(Hu et al., 2018) conducted research on several approaches to intrusion detection in ICS settings. Protocol analysis-based intrusion detection, traffic mining-based intrusion detection, and control process analysis-based intrusion detection are all explored. ICS protocol analysis identifies malicious behaviour by creating a baseline of an ICS protocol's usual behaviour in a particular environment and reporting any deviations from the baseline. Traffic mining-based intrusion detection takes traffic data from the ICS environment and detects abnormalities using data mining and analysis methodologies. Control process-based intrusion detection detects abnormalities by collecting data from an industrial process, including all its components. The suggested techniques enable network monitoring between ICS systems or sites as well as between the OT and IT networks. The "Shadow Security Unit (SSU)" is a security monitoring system proposed by (Cruz et al., 2015) that can intercept communications from a PLC or RTU to continually monitor and identify irregularities. The proposed device keeps track of any correspondence between a PLC/RTU and the master station. Figure 3 depicts the SSU's deployment.

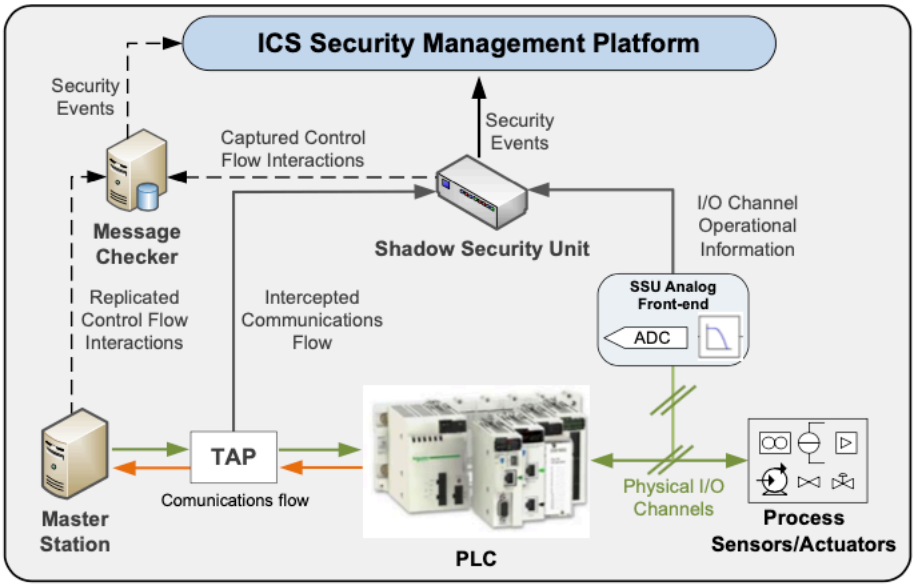


Figure 3: Deployment of the "Shadow Security Unit" (Cruz et al., 2015)

The SSU is installed in tandem with the PLC/RTU, allowing monitoring of both the communication flow to the master station and the physical I/O channels to field devices. The SSU's simultaneous deployment ensures that it does not disrupt the normal operation of the PLC or the overall system. The SSU monitors and detects threats from physical devices to the PLC and, ultimately, the master station.

5.4 Secure Remote Access

The convergence of IT and OT has resulted in ICS systems being exposed over the internet. The efficiencies brought about include the remote management and operation of ICS systems. The same channels utilised to remotely access OT environments can, however, be maliciously used by threat actors to compromise the systems in these OT settings. Secure remote access should be essential not just for ICS operators but also for vendors and third-party operators (Lee and Conway, 2022).

Bruzgiene and Jurgilas (2021) present a Multi-Factor Authentication technique for remote access to ICS systems worth exploring. The procedure begins with a user submitting their credentials. If correct, an authentication request will be issued to their mobile device. When the user confirms the authentication request, a permission request is issued to the user's supervisor. The supervisor must assess whether the individual user is permitted to connect remotely and whether they have the required rights to access the system in question. Only if a supervisor approves the permission request will the user be granted access to the ICS system. Figure 4 depicts this procedure.

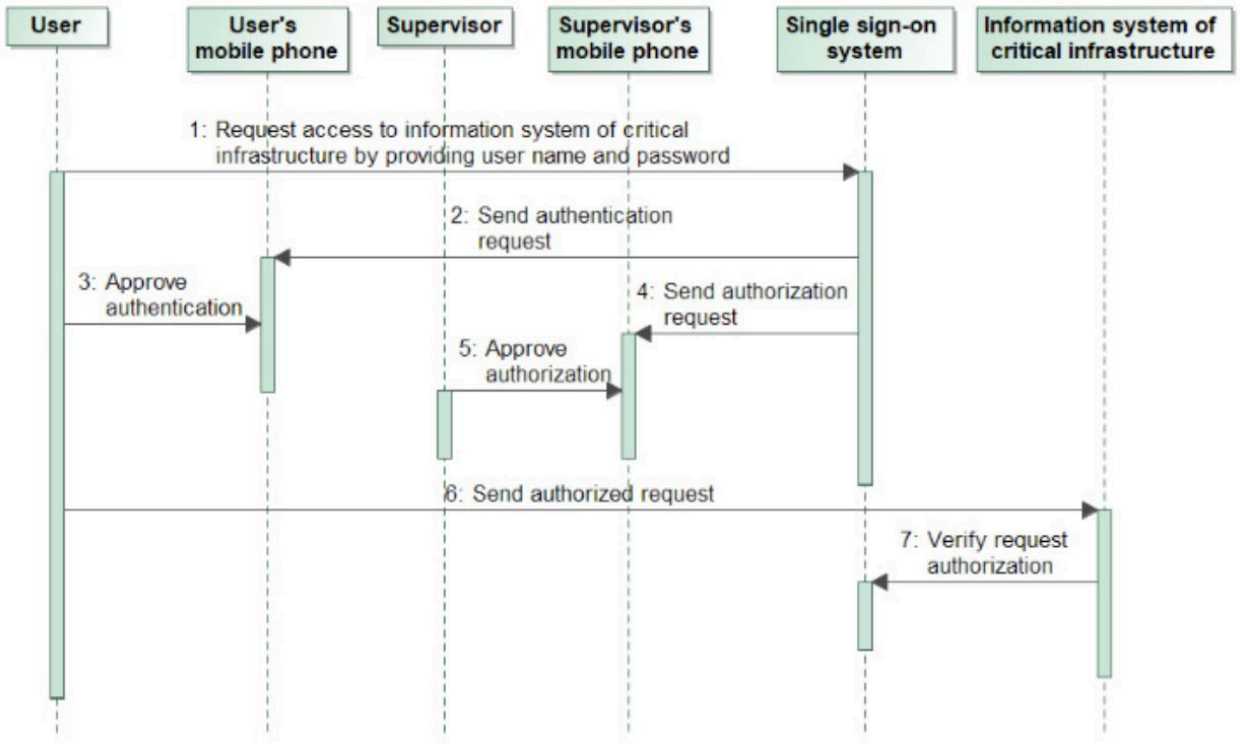


Figure 4: Proposed MFA Method (Bruzgiene and Jurgilas, 2021)

Although such an elaborate access control approach would be hard to compromise, scalability concerns are raised. What is the suggested supervisor-to-user ratio for such a system? Will the absence of supervisors affect the functionality of critical infrastructure? Secure access control is crucial in restricting access to the ICS by threat actors.

5.5 Risk-Based Vulnerability Management

Vulnerability management involves the identification, analysis, prioritisation, and remediation of vulnerabilities (Ndichu et al., 2020). In contrast to IT, where security updates are prioritised to safeguard sensitive data, OT settings have quite different approaches to managing vulnerabilities. OT gives the consistent performance of industrial processes a higher priority than security updates that could impair such operations. Because just 4% of identified ICS vulnerabilities require immediate action (Lee and Conway, 2022), OT systems can spend several years without receiving security updates. OT vulnerability management should be centred around avoiding process disruptions and prioritising vulnerabilities that grant threat actors access to the ICS, addressing vulnerabilities being actively exploited and those that introduce functionality that can be abused, especially to achieve scenarios identified in incident response planning (Lee and Conway, 2022).

To determine the significance and criticality of ICS vulnerabilities, (Ani, He, and Tiwari, 2020) offer a probabilistic approach. The "Multi-Attribute Vulnerability Criticality Analysis (MAVCA) model" takes three things into account: 1) the severity of the vulnerability, 2) the probability of the vulnerability being exploited, and 3) the functional dependencies of a vulnerable host. This allows for environment-specific prioritisation of vulnerabilities. An expansion of the Common Vulnerability Scoring System (CVSS), which is used to assess the criticality of vulnerabilities, is proposed by (Ur-Rehman et al., 2020). The "$CVSS_{IoT\text{-}ICS}$ Framework" expands CVSS to take into account variables specific to ICS settings, as depicted in Figure 5.

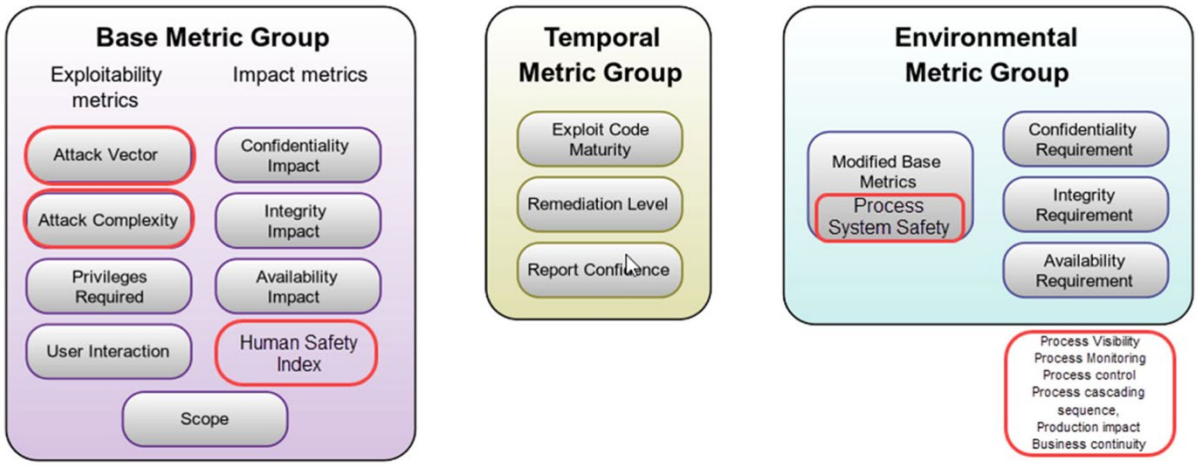


Figure 5: The $CVSS_{IoT-ICS}$ Framework (Ani, He, and Tiwari, 2020)

6. Mapping ICS Critical Controls to NERC CIP Standards

The North American Electric Reliability Corporation (NERC) developed a set of standards for the protection of critical infrastructure, also known as NERC CIP Standards. The standards provide a framework to manage the security of the Bulk Electric System (BES) and associated critical assets in North America (Francia and El-Sheikh, 2022). This section maps the NERC CIP Standards and their associated requirements to the SANS ICS Critical Controls.

The NERC CIP-008 and CIP-009 standards cover incident response and disaster recovery, which map to the first control of the SANS ICS Critical Controls - ICS specific indent response. The NERC CIP-008 provides guidelines on establishing a cybersecurity incident response plan, including documenting incident handling procedures and personnel responsible for carrying out the incident response plan. The NERC CIP-008 standard advises that incident response plans must be tested at least once every 15 calendar months. This is to ensure that the incident response plan is sufficient for responding to the said incident and provides an opportunity for the personnel responsible for incident response to improve their processes. NERC CIP-008 emphasizes the importance of documenting lessons learned from each incident, or tabletop exercises. Members of the incident response team are to be notified no later than 90 calendar days of any changes to the incident response plan. The NERC CIP-008 ensures coordination and consistency in responding to cyber incidents. Also mapped to ICS incident response is NERC CIP-009, which provides guidelines on establishing disaster recovery plans. This is crucial as some incidents, such as ransomware, may necessitate recovery from backups, making disaster recovery important.

Mapped to defensible architecture is the first requirement of NERC CIP-005, which provides guidelines on establishing electronic security perimeters. NERC CIP-005 requires that assets reside within defined electronic security perimeters and the establishment of methods for the detection of malicious communications. ICS network visibility and monitoring are mapped to NERC CIP-005, NERC CIP-007, and NERC CIP-012. These standards are centred around providing network visibility, including monitoring communications and security events. Requirements 2 and 3 of NERC CIP-005 provide guidelines for implementing remote access to ICS. Requirement 2 focuses on interactive remote access, mandating the use of multi-factor authentication. Also recommended is the use of an intermediate system, such that the initiating system does not directly access the ICS. Requirement 3 provides guidelines on managing vendor remote access, including having controls in place to determine if a vendor has initiated a remote access session into an environment and methods to terminate the session when necessary. Table 1 presents the mapping of ICS Critical Controls to NERC CIP standards and associated requirements.

Table 1: A mapping of ICS critical controls to NERC CIP Standards

ICS Critical Controls	NERC CIP
ICS Incident Response	CIP-008-6: Incident Reporting and Response Planning R1: Cybersecurity Incident Response Plan R2: Cybersecurity Incident Response Plan Implementation and Testing R3: Cyber Security Incident Response Plan Review, Update, and Communication R4: Notifications and Reporting for Cyber Security Incidents

ICS Critical Controls	NERC CIP
	CIP-009-6: Recovery Plans for BES Cyber Systems R1: Recovery Plan Specifications R2: Recovery Plan Implementation and Testing R3: Recovery Plan Review, Update and Communication
Defensible Architecture	CIP-005-7: Electronic Security Perimeter(s) R1: Electronic Security Perimeter
ICS Network Visibility and Monitoring	CIP-005-7: Electronic Security Perimeter(s) R1: Electronic Security Perimeter CIP-007-6: Systems Security Management R1: Ports and Services R3: Malicious Code Prevention R4: Security Event Monitoring CIP-012-1: Communications between Control Centers R1: Securing and Monitoring Communications
Secure Remote Access	CIP-005-7: Electronic Security Perimeter(s) R2: Remote Access Management R3: Vendor Remote Access Management for EACMS and PACS
Risk-Based Vulnerability Management	CIP-007-6: Systems Security Management R2: Security Patch Management CIP-010-4: Configuration Change Management and Vulnerability Assessments R1: Configuration Change Management R2: Configuration Monitoring R3: Vulnerability Assessments

Vulnerability management is mapped to NERC CIP-007 and NERC CIP-010. NERC CIP-007 provides guidelines on managing security updates, including evaluating every 35 calendar days security updates applicable to assets in the OT environment. NERC CIP-010 provides guidelines on securely handling system configurations, avoiding exposing systems to attacks. The standard also provides guidelines on managing vulnerabilities, including conducting vulnerability assessments every 15 calendar months, both in test and production environments. This section maps the ICS Critical Controls to a set of standards defined by NERC. The next section concludes the paper.

7. Conclusion

The objective of this paper was to examine the current literature on ICS security, with a particular focus on identifying security controls that assists ICS operators in protecting their systems in compliance with the SANS Five ICS Cybersecurity Critical Controls. Industrial organizations are to establish ICS-specific incident response procedures, taking into consideration common attacks in their sector. This leads to the implementation of a network architecture that facilitates the collection of data to support the incident response process whilst satisfying operational requirements. Such a network should allow for visibility into communications between systems, including OT and IT environments and devices-to-device communications. Organizations should implement multi-factor authentication mechanisms to ensure secure remote access for both operators and vendors. Finally, there should be a prioritization of vulnerabilities, with those that introduce functionality to the ICS being remediated. The paper also mapped the ICS critical controls against the NERC CIP standards. Future work should expand on this effort by mapping the controls to other security standards in ICS.

References

Alladi, T., Chamola, V. and Zeadally, S., 2020. Industrial control systems: Cyberattack trends and countermeasures. Computer Communications, 155, pp.1-8.

Ani, U.D., He, H. and Tiwari, A., 2020, June. Vulnerability-based impact criticality estimation for industrial control systems. In 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security) (pp. 1-8). IEEE.

Asghar, M.R., Hu, Q. and Zeadally, S., 2019. Cybersecurity in industrial control systems: Issues, technologies, and challenges. Computer Networks, 165, p.106946.

Bolton, W., 2021. Instrumentation and control systems. Newnes.

Bruzgiene, R. and Jurgilas, K., 2021. Securing remote access to information systems of critical infrastructure using two-factor authentication. Electronics, 10(15), p.1819.

Cruz, T., Barrigas, J., Proença, J., Graziano, A., Panzieri, S., Lev, L. and Simões, P., 2015, May. Improving network security monitoring for industrial control systems. In 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM) (pp. 878-881). IEEE.

Drias, Z., Serhrouchni, A. and Vogel, O., 2015, August. Analysis of cyber security for industrial control systems. In 2015 international conference on cyber security of smart cities, industrial control system and communications (ssic) (pp. 1-8). IEEE.

Drias, Z., Serhrouchni, A. and Vogel, O., 2015, August. Analysis of cyber security for industrial control systems. In 2015 international conference on cyber security of smart cities, industrial control system and communications (ssic) (pp. 1-8). IEEE.

Firoozjaei, M.D., Mahmoudyar, N., Baseri, Y. and Ghorbani, A.A., 2022. An evaluation framework for industrial control system cyber incidents. International Journal of Critical Infrastructure Protection, 36, p.100487.

Francia III, G.A. and El-Sheikh, E., 2022. NERC CIP Standards: Review, Compliance, and Training. Global Perspectives on Information Security Regulations: Compliance, Controls, and Assurance, pp.48-71.

Hahn, A., 2016. Operational technology and information technology in industrial control systems. In Cyber-security of SCADA and other industrial control systems (pp. 51-68). Springer, Cham.

Hemsley, K. and Fisher, R., 2018, March. A history of cyber incidents and threats involving industrial control systems. In International Conference on Critical Infrastructure Protection (pp. 215-242). Springer, Cham.

Hu, Y., Yang, A., Li, H., Sun, Y. and Sun, L., 2018. A survey of intrusion detection on industrial control systems. International Journal of Distributed Sensor Networks, 14(8), p.1550147718794615.

Lee, R.M., and Conway, T., 2022. The Five ICS Cybersecurity Critical Controls. Available from: https://www.sans.org/white-papers/five-ics-cybersecurity-critical-controls/ (Accessed 5 January 2023)

Miller, T., Staves, A., Maesschalck, S., Sturdee, M. and Green, B., 2021. Looking back to look forward: Lessons learnt from cyber-attacks on Industrial Control Systems. International Journal of Critical Infrastructure Protection, 35, p.100464.

Murray, G., Johnstone, M.N. and Valli, C., 2017. The convergence of IT and OT in critical infrastructure.

Ndichu, S., McOyowo, S., Okoyo, H. and Wekesa, C., 2020. A Remote Access Security Model based on Vulnerability Management. Int. J. Inf. Technol. Comput. Sci, 12(5), pp.38-51.

Piggin, R., 2014. Industrial systems: cyber-security's new battlefront [Information Technology Operational Technology]. Engineering & Technology, 9(8), pp.70-74.

Pliatsios, D., Sarigiannidis, P., Lagkas, T. and Sarigiannidis, A.G., 2020. A survey on SCADA systems: secure protocols, incidents, threats and tactics. IEEE Communications Surveys & Tutorials, 22(3), pp.1942-1976.

Staves, A., Balderstone, H., Green, B., Gouglidis, A. and Hutchison, D., 2020, May. A framework to support ICS cyber incident response and recovery. In the 17th International Conference on Information Systems for Crisis Response and Management.

Stouffer, K., Falco, J. and Scarfone, K., 2011. Guide to industrial control systems (ICS) security. NIST special publication, 800(82), pp.16-16.

Stouffer, K., Pease, M., Tang, C., Zimmerman, T., Pillitteri, V. and Lightman, S., 2022. Guide to Operational Technology (OT) Security (No. NIST Special Publication (SP) 800-82 Rev. 3 (Draft)). National Institute of Standards and Technology.

Thomas, R.J. and Chothia, T., 2020. Learning from Vulnerabilities-Categorising, Understanding and Detecting Weaknesses in Industrial Control Systems. In Computer Security (pp. 100-116). Springer, Cham.

Ur-Rehman, A., Gondal, I., Kamruzzaman, J. and Jolfaei, A., 2020. Vulnerability modelling for hybrid industrial control system networks. Journal of Grid Computing, 18(4), pp.863-878.

Yadav, G. and Paul, K., 2021. Architecture and security of SCADA systems: A review. International Journal of Critical Infrastructure Protection, 34, p.100433.

Zhou, C., Hu, B., Shi, Y., Tian, Y.C., Li, X. and Zhao, Y., 2020. A unified architectural approach for cyberattack-resilient industrial control systems. Proceedings of the IEEE, 109(4), pp.517-541.

NCSS: A Global Census of National Positions on Conflict, Neutrality and Cooperation

Radu Antonio Serrano Iova[1] and Tomoe Watashiba[2]
[1]Ragnar Nurkse Department, School of Business and Governance, Tallinn University of Technology, Tallinn, Estonia
[2]School of Governance, Law and Society, Tallinn University, Tallinn, Estonia

raduantonio.serranoiova@taltech.ee
tomoetw@tlu.ee

Abstract: The ubiquity of ICT and the increase in cyber threats have pushed countries to view cybersecurity from a national perspective and draft appropriate national strategies on the topic. While containing similar terminology, these strategies are tailored to the national contexts and hence, differ across regions, cultures, and political contexts. Previous research of these documents has been focused on comparative analysis of countries that can either be considered well developed on this topic or for specific subtopics of cybersecurity. However, some of the subtopics have not been addressed, only now having become more prevalent due to current international conflicts and national / regional socio-political scuffles that have spilled into cyberspace. In our paper, we investigate all countries that have published a National Cyber Security Strategy - NCSS - (or any similar document under a different nomenclature, e.g., policy, decree, etc.), specifically in reference to their position on war, neutrality, and international cooperation. Countries maintaining an NCSS will first be identified using international databases, upon which further study of the aforementioned topics in the NCSSs will occur. We hypothesize, that while international cooperation will be present in most, if not all NCSSs, armed conflicts and neutrality will not be addressed at all nor in depth, in those that contain any reference to them. The resulting paper will present a near-global case study of these topics, which can then signify potential areas of improvement, capacity building, and strengthening of democratic coalitions, globally.

Keywords: National Cyber Security Strategy, International Cooperation, Neutrality, Conflicts, Cyberspace

1. Introduction

Russia's latest invasion of Ukraine was not only undertaken through physical means. Since 2014, the latter has been on the receiving end of targeted cyber attacks which only increased during the invasion. Satellite networks, border checkpoints, telecommunication service providers, banks, power stations, governmental services, and even non-governmental, charity, and aid organizations, among others, were attacked in the first two months of the conflict (Przetacznik and Tarpova 2022). Additionally, in 2022 only, Andorra, Australia, Bahrain, China, Costa Rica, Ethiopia, most EU countries (and its own Commission), India, Iran, Jordan, the Marshall Islands, Mexico, Pakistan, the U.K., the U.S.A., Vanuatu, and many other countries, have been targets (and sometimes even perpetrators) of cyber-attacks. Not even international humanitarian organizations, such as the International Committee of the Red Cross, have been exempted from such aggressions (CSIS no date).

These examples and the ubiquity of ICT have resulted in countries publishing their own National Cyber Security Strategies (NCSSs), or similar documents. Although these documents differ in their formatting characteristics, they all present their nation's official position on the issue of cybersecurity. Because of the invasion and the never-ending number of cyber-attacks, we decided to investigate all the countries' references on war, neutrality and international cooperation in these documents. We hypothesize that very few United Nations' (U.N.'s) members have declared their cybersecurity actions during war nor discussed neutrality in any way in their current NCSS. Moreover, we believe that all NCSSs will talk about international cooperation, given the borderless nature of cyberspace, and that countries will reference their allied preferences. This resulting paper will serve as a snapshot of the topics, allowing individual countries to improve themselves and their bi- and multilateral endeavors.

2. Strategy, topics and previous studies

Prior to anything, the concept of a National Cyber Security Strategy (NCSS), or similar document, should be defined. Traditionally, "[a] *strategy of a business forms a comprehensive master approach that states how the business will achieve its mission and objectives"* (Wheelen et al. 2017, p.50). Transposed to the public sector and scaled up to a larger context, a national strategy is *"the identification of national interests and ambitions and the use of various resources (national and other) to preserve or pursue those interests and ambitions"* (Cornish, Lindey-French and Yorke 2011). Framed more specifically toward the notion of the security of a country, it becomes then a national security strategy. These documents are usually addressed to "*adversaries or potential*

enemies" (Drew and Snow 1988, p.48). However, by being public, in addition to serving as possible intimidation to an enemy, they might serve political purposes such as influencing public opinion, swaying specific constituencies (Caudle 2009), and even communicating with existing and potential allies, or non-aligned stakeholders. The cyberspace has become part of national security, as evidenced by multiple writings (e.g. Reveron 2012; Yannakogeorgos and Lowther 2016; Libicki 2018) throughout the years. Therefore, an NCSS is a document that identifies interests and ambitions related to national cybersecurity and communicates the use of resources to reach those goals.

Like national security strategy, an NCSS can "*vary widely in length, format and complexity*" (DuMont 2019), from country to country. Given that language, year of publication, and national contexts differ across the world's NCSSs, this variety can easily affect the title and contents of the document. However, as the logical offshoot of the former, some of an NCSS' characteristics are handed down. Some of these, compiled by DuMont (2019) include: having an endorsement by the governmental leader, reflecting national values, articulating national interests, declaring a strategic vision, identifying and assessing future challenges, and containing risk assessments, resources overviews, timeframes effectiveness measurements and implementation guidance. Therefore, a high-level document that presents most, if not all, of these elements would be considered as an NCSS and studied as one of its peers. Some outlier variations have been named 'national policy' or 'national doctrine'

Another variable that might have also affected the selection process is the 'cyber security' term. Sidestepping the linguistical connotations of having it in a single word, or not, some documents are titled using the terms information security, digital security, and Information and Communications Technology (ICT) security. While these were sporadic cases, and some even due to their translation into English or original date of publication, we have taken them all as synonyms. Von Solms and Van Niekerk's (2013) position on two of those terms (even though they acknowledge that cyber security and information security are often used interchangeably) could shed additional light for those inquiring about the similarities and differences in terminology. Nevertheless, it is within this variety of terminology and contexts that we felt that a global study was necessary to discover and present the countries' positions on the topic of warfare, neutrality, and cooperation.

2.1 Warfare / War/ Conflict

According to Merriam-Webster's dictionary, 'warfare' is defined as "*military operations between enemies (hostilities, war),*" also meaning "*an activity undertaken by a political unit (such as a nation) to weaken or destroy another.*" (Merriam-Webster no date) Oppenheim's International Law similarly defines war as "*a contention between two or more states, through their armed forces, for the purpose of overpowering each other and imposing such conditions of peace as the victor pleases*" (Lauterpach 1952, p.220). In understanding the use of 'warfare' or 'war' in our analysis, we have followed these general definitions, essentially understanding war or warfare as a hostile operation between two or more parties, usually involving extensive employment of forces aimed at countering the other(s). Conflict is a state where friction exists between parties, diplomatically, economically, politically, militarily and/or otherwise; yet to treat some conflicts as 'war' for states has significant legal implications and consequences, such as the application of the law of war and the law of neutrality, suspension of non-hostile relations (e.g. trade, diplomatic, treaty relations) between the belligerents, as well as the prohibition of the Use of Force under Article 2(4) of the U.N. Charter (Greenwood 1987). Therefore, in a factual sense, States avoid using 'state of war,' yet often, in the colloquial use of the terms, 'war', 'warfare' or 'conflict' are used more or less interchangeably. In this analysis, we, therefore, have followed the interchangeable usage of these three words.

2.2 Neutrality

In this analysis, the word 'neutrality' has been identified from two different angles, one from the law of neutrality as defined in the International Humanitarian Law (IHL) and the other from the internet/technological neutrality. The law of neutrality, when applied in cyberspace, entails duties such as abstention/non-participation, better expressed as "…*abstain*[ing] *from committing any acts of kinetic or cyber hostility against belligerents and providing them with military assistance, such as the provision of cyber weaponry or the recruitment of a cyber "corps of combatants"…"* or prevention, meaning *"…neither allow nor tolerate certain types of malicious activities on its territory and infrastructure*…" as well as the maintenance of impartiality, "…*apply[ing] every restricting measure and prohibition in the context of its neutral duties and rights in a non-discriminatory manner toward all belligerents.*" (Cordey and Kohler 2021, p.1). There are other elements, including the respect for a neutral state's territorial integrity and the ban of cyber operations against/from/through neutral nations,

particularly for a belligerent country, with both neutral and belligerent states to be taken into account. The law of neutrality's applicability is confined to international armed conflicts (IAC) as defined in IHL and state actors.

Internet / technological neutrality is another aspect of neutrality that is relevant to this analysis, and it is generally understood to mean that Internet Service Providers (ISPs) ensure 'neutrality' in the flow of the internet traffic, treating equally, the contents that pass through their cables, cellular towers, satellites and other core infrastructures (Wu 2003), and "[w]*hen providing internet access services, providers of those services should treat all traffic equally, without discrimination, restriction or interference, independently of its sender or receiver, content, application or service, or terminal equipment*" (European Union 2015, Preamble (8)).

2.3 International Cooperation

In this analysis, we have approached the term 'international cooperation' from the perspectives of both development assistance with a closer connotation to capacity-building and part of confidence-building measures where information and intelligence sharing may be conducted between/among countries. The capacity-building, in general term, is defined by the U.N. as "*the process of developing and strengthening the skills, instincts, abilities, processes and resources that organizations and communities need to survive, adapt, and thrive in a fast-changing world*" (U.N. no date) and in a more specific case of 'cyber capacity-building,' we have followed the meaning of assistance or cooperation to "*strengthen a country's legal, technical, and policy capability, and protect against malicious cyber activity*" (Naylor, Painter and Hakmeh 2022). Confidence-Building Measures (CBMs) are considered to be "*a verified instrument of international politics, which aims to prevent the outbreak of war or an (international) armed conflict by miscalculation or misperception of the risk, and the consequent inappropriate escalation of a crisis situation*" and they "*achieve this by establishing practical measures and processes for (preventive) crisis management between States.*" (Ziolkowski 2013, p.5). In the cyberspace context, CBMs could include sharing, providing, and exchanging information about cyber threats and vulnerabilities as well as best practices, facilitating communications through voluntary mechanisms such as the meeting of experts, provision of contact data, sharing of information through workshops, etc., at international and regional levels.

2.4 Past studies

Previous studies regarding NCSSs have compared EU and NATO cybersecurity strategies to national cyber security strategies (Štitilis, Pakutinskas and Malinauskaitė2017), attempted to discover governance trends (Shackelford and Kastelic 2014) and even present correlations between NCSS development to that of a digital economy (Teoh and Mahmood 2017). Other studies, for example, have focused on a comparative contents analysis of NCSSs of selected countries, often considered to be more 'advanced' or 'elaborate,' to identify differences and similarities, and to draw conclusions on recommended approaches and contents to be potentially considered by policymakers (OECD 2012; Luiijf, Besseling and Graaf 2013; Newmeyer 2015; Sabillon, Cavaller and Cano 2016; Shafqat and Masood 2016). In a similar vein, an attempt to utilize the quantitative analytical tool (e.g., Latent Dirichlet Allocation) for more comprehensive and automated analysis and understanding of texts has also been conducted (Kolini and Janczewski 2017), as well as more in-depth country-specific investigations (for example, see Daricili and Özdal 2018; Hurel 2020; Beecroft 2021).

3. Methodology

The methodology used for this article involved the systematic collection and synthesis of existing information, i.e. literature review. A literature review is a great way to summarize findings to uncover areas where research (Snyder 2019) or improvements are needed. As described by Snyder (2019), the aim of a systematic review allows for the identification of all evidence that fits a criterion to answer a particular question or hypothesis.
Out of the 193 member states of the United Nations, this study was conducted on the states that had publicly available National Cyber Security Strategies or similarly high-level national documents on the topic, in December 2022 and early January 2023. Identification of these states was conducted mainly through the eGA's National Cyber Security Index (available at https://ncsi.ega.ee/) and UNIDIR's Cyber Policy Portal (available at: https://cyberpolicyportal.org/). The former is *"… a global live index, which measures the preparedness of countries to prevent cyber threats and manage cyber incidents …"* and *"… also a database with publicly available evidence materials and a tool for national cyber security capacity building."* (eGA no date) The latter is *"an interactive map of the global cyber policy landscape. It provides profiles of the cyber policies of all 193 UN Member States…"* and *"… seeks to support informed participation by relevant stakeholders in all policy processes*

and promote trust, transparency, and cooperation in cyberspace." (UNIDIR no date) The benefit of these tools is that they provide direct links to official third-party websites that might contain the NCSS or similar document. The NCSS, of every country that had one, was checked to verify that it was an official document (i.e., published by official sources, the government of that country) and that it was the current version. For this latter characteristic, even if a NCSS had an expiry date, it would still be considered the current one if it was still publicly present in a governmental website and if no other NCSS had been published. For example, there was a nation that had a previous version of their NCSS in English, but the current one was only found in their official language. In this case, the one in the official language was the current official document.

After all corresponding NCSSs were identified, they were individually searched on the three proposed topics: warfare, neutrality and international cooperation. The keywords used were 'war', 'warfare', 'conflict', for the first topic, 'neutrality', 'neutral', for the second one, and 'cooperation', 'collaboration' in the international context, for the third one, with the corresponding terms in other languages. Their presence was noted, grouped, and analyzed further, as presented in the next section. For a more compact presentation of the results, the countries presented in the subsequent section have been noted using the ISO 3166-1 alpha-3 country codes, in the alphabetical order of their full names. The ISO 3166 standard can be found at: https://www.iso.org/iso-3166-country-codes.html

4. Results

In the end, 113 (out of the 194) U.N. member states had publicly accessible current NCSS, or a similar high-level document. Most NCSSs were presented in their national languages, but they also had a version in English, which allowed for an easier execution the topic search. After English, the NCSSs were found to be drafted in Spanish, French, Russian, Portuguese, Arabic and Romanian, with the remaining ones each in a different language (see

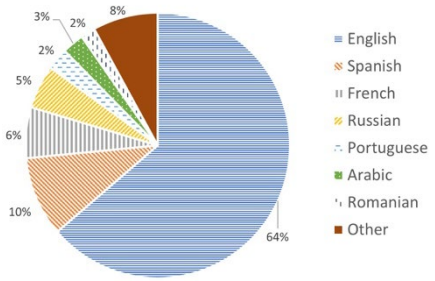


Figure 1 Language availability of the NCSSs

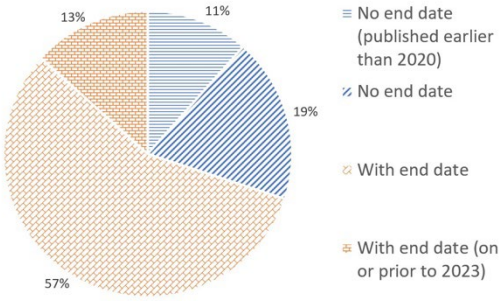


Figure 2 NCSSs with and without end dates). While there is no obligation to publish an NCSS in English or any alternate language, the majority of the countries have done so, with some even specifying that the translation is a courtesy and should any discrepancies arise, the NCSS in their national language shall prevail.

The large majority of the NCSSs have been established for a definite period of time, i.e., 79 of them contain an end date. However, 15 of them have an end date of 2023 or prior, meaning that they are close to or already expired. The remaining 34 NCSSs do not have an end date, out of which 13 have been published prior to 2020, with the oldest entry dating back to 2003, followed by 2011.

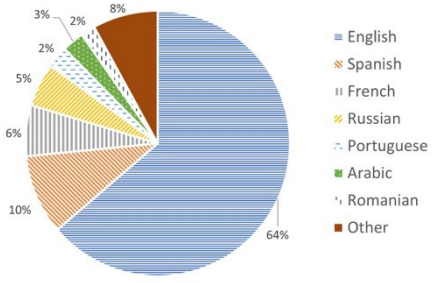


Figure 1 Language availability of the NCSSs

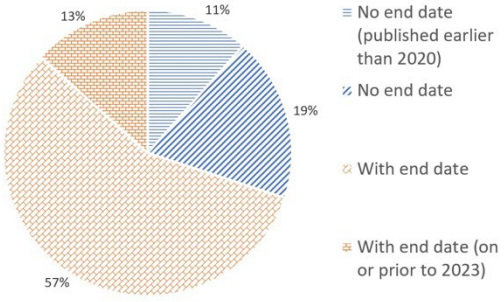


Figure 2 NCSSs with and without end dates

On the topic of warfare, 66 of the NCSS mention the terms 'war', 'warfare', 'conflict' and their types, in multiple lines of context, sometimes more than once in a single document (Figure 3). More than half of them allude to war / conflict, or a specific type thereof, as a threat to the nation, and in most cases in the sections describing the current situation of the country or the reasons behind the creation of the NCSS. As presented in Figure 3, in this context, we have the highest variety of descriptors referencing different types of warfare or conflict. The

second most common context was regarding the prevention of war or conflicts. Similarly, there was not much specificity, with the common trend being the need to prevent such armed or cyber actions. Next, we have thirteen nations that talk about their preparations for or existing preparedness in case cyber, hybrid, armed or cyberspace warfare. These will be discussed in more detail in upcoming paragraphs describing Figure 4. In the less mentioned contexts, six countries clearly state the entity / organization responsible for cyber warfare, while four of them outright express that it is an area of improvement for their state, and three others declare that the cyberspace is a domain of warfare. Finally, in the miscellaneous context we have both Japan and Malta, since the former alludes to the necessity of international collaboration, while the latter that it would work to establish its own national position.

Type of War / Warfare / Conflict	Context	Countries
Cyber, Hybrid, Military, Information, Cyberspace, International, Armed, Asymmetric, Electronic, Psychological, Imagological, Command-control	As a threat	AUS - AUT - BEL - BGR - BFA - CYP - EGY - EST - ETH - GEO - GRC - HUN - IRQ - JPN - KEN - LTU - LUX - MLT - MCO - MNE - MOZ - NLD - NGA - PNG - PHL - PRT - KOR - MDA - RWA - SRB - SVN - CHE - TJK - TKM - GBR
Cyber	As a domain of Warfare	ALB - LTU - UKR
Information, Cyber, Electronic	As area of Improvement	GMB - LBN - ZAF - TJK
Cyber	Organizational Responsibilities	BRA - BGR - SWZ - SEN - ZMB - VNM
Armed, Cyber, Cyberspace	Prevention of	ARG - AUT - CHL - CHN - HRV - CZE - FRA - DEU - MCO- NZL - ROU - ESP - UGA - USA
Cyber, Hybrid, Cyberspace, Armed	Preparedness	DEU - LVA - MUS - NLD - MKD - NOR - POL - KOR - RUS - ZAF - SWE - CHE - THA
Cyberspace, International	Miscellaneous	JPN - MLT

Figure 3 Mentions of war / warfare / conflict in 66 NCSSs

Figure 4 presents the 13 countries that reference their preparations and preparedness for or during times of war. Most of them allude to existing capabilities, albeit in extremely short descriptions, while a few of them state that they will or must prepare such capabilities. There is no other presential common trend apart from the fact that a majority of these NCSS are currently outside their ending date.

Country	Title of the Document	Preparedness
DEU	Cyber Security Strategy for Germany 2021	In case of conflict, contact persons will be available, and already established, reliable channels of communication can be used. (p. 114)
LVA	Informatīvais ziņojums "Latvijas kiberdrošības stratēģija 2019.–2022. gadam" (Informative report "Latvia's Cyber Security Strategy for 2019-2022")	Similarly, in times of crisis and war, the government must ensure the protection of information and cyberspace using active and passive defence measures to prevent external influence on the population and paralyzing government action (Task 2.2). (p. 17)
MUS	Republic of Mauritius National Cyber Security Strategy 2014 - 2019	National cyber resilience will be tailored to ensure the preparedness and predictive capabilities required by the goals and to facilitate its operating capability during cyber conflicts and post-conflict recovery. (p. 8)
NLD	Nederlandse Cybersecuritystrategie 2022-2028 (Dutch Cyber Security Strategy 2022-2028)	The government has offensive and defensive cyber capabilities that are effective in times of peace and war. (p. 38)
MKD	Republic of Macedonia National Cyber Security Strategy 2018 - 2022	5. Development of national procedures in time of peace, crisis, a state of emergency and state of war, in order to manage incidents which will enable efficient intra-institutional cooperation, where every institution have a pre-defined role, will employ appropriate protocols and procedures, as well as information exchange, communication and coordination channels. (p. 18)
NOR	National Cyber Security Strategy for Norway	Civilian support of the Norwegian Armed Forces in the event of cyber security challenges in times of crisis and armed conflict is provided within the framework of the total defence concept. (p. 10)
POL	Cybersecurity Strategy of the Republic of Poland for 2019 - 2024	The Armed Forces of the Republic of Poland, as the fundamental element of the state defence system, should be involved in activities in cyberspace at the same level as in the air, on the ground and at sea, in peacetime, during war and in crisis situations alike. (p. 24)
KOR	National Cybersecurity Strategy	Promote cooperation in sectors such as national defence, intelligence, and law enforcement, as well as exchange with the private sector to respond to cybersecurity threats, including acts of war, terrorism, and crime. (p. 23)
RUS	Указ Президента Российской Федерации от 05.12.2016 г. № 646 Об утверждении Доктрины информационной безопасности Российской Федерации (Decree of the President of the Russian Federation of December 5, 2016 No. 646 On approval of the Information Security Doctrine of the Russian Federation)	8. National interests in the information sphere are: ... ...b) ensuring the stable and uninterrupted functioning of the information infrastructure, primarily the critical information infrastructure of the Russian Federation (hereinafter referred to as the critical information infrastructure) and the unified telecommunication network of the Russian Federation, in peacetime, during the immediate threat of aggression and in wartime;
ZAF	National Cybersecurity Policy Framework for South Africa	In order to protect its interests in the event of a cyber-war, a cyber defence capacity has to be built. (p. 24)
SWE	A national cyber security strategy Skr. 2016/17:213	An effective national cyber defence is developed and strengthened in peacetime and as part of total defence planning and must be capable of functioning in peace, crisis and war. (p. 18)
CHE	National strategy for the protection of Switzerland against cyber risks (NCS) 2018-2022	In order to prevent such activities, Switzerland must therefore include cyber defence and cyber diplomacy in its preparations for potential conflict. (p. 4 - 5)
THA	National Cybersecurity Strategy 2017-2021	3. Have a plan to face cyber threats when there is a national cyber crisis or cyber war. (p. 39)

Figure 4 Preparedness in case of war / warfare / conflict

Out of the 133 NCSS, only five countries reference the concept of neutrality in their documents. Two of them refer to Internet neutrality, two others to technical/technological neutrality, and the last one to Information neutrality (Figure 5). No other NCSS explores the any neutrality concepts, associated or not, with cybersecurity.

Country	Title of the Document	Mention of Neutrality
BLR	Концепция информационной безопасности Республики Беларусь (Concept of Information Security of the Republic of Belarus)	CHAPTER 8 **INFORMATION NEUTRALITY** 31. In international relations, the information sovereignty of the Republic of Belarus is ensured, among other things, on the basis of the principle of **information neutrality**, which provides for a peaceful foreign information policy, respect for the generally recognized and generally accepted rights of any state in this area, and the exclusion of the initiative to interfere in the information sphere of other countries aimed at discrediting or challenging their political, economic, social and spiritual standards and priorities, as well as damaging the information infrastructure of any states and participating in their information confrontation...
CHL	National Cybersecurity Policy	Therefore, the policy includes and promotes the following: ... - ...the principle of respecting **Internet neutrality** is also included, so that Internet service providers may not discriminate or arbitrarily restrict the access to any content whatsoever, unless there is a legal justification to do that. (p. 20)
SWZ	Eswatini National Cybersecurity Strategy 2020 - 2025	It is important that the established cybersecurity legal and regulatory framework for Eswatini is suitably applicable and **technology neutral**. (p. 21)
KGZ	Стратегия кибербезопасности Кыргызской Республики на 2019-2023 годы (Cybersecurity Strategy of the Kyrgyz Republic for 2019-2023)	At the same time, it is necessary to take into account international experience, best practices and recommendations, but at the same time proceed from the specific conditions of the Kyrgyz Republic, which include the following: ... - an extremely small volume of the domestic market of tools and solutions for the information technology and communications industry, including the cybersecurity sector of the Kyrgyz Republic, and almost complete dependence on foreign suppliers of software and hardware products. This circumstance increases the importance of the task of developing a national certification system for imported products in the field of information technology, as well as testing it for vulnerability and undeclared capabilities in order to maintain **technical neutrality** and sovereignty.
ESP	National Cybersecurity Strategy 2019	To do this, it [Spain] will defend an interoperative, **neutral**, open and diverse **internet**, a reflection of international cultural and linguistic plurality, based on a system of democratic, representative and inclusive governance resulting from agreement and consensus. (p. 39)

Figure 5 Mentions of neutrality in only 5 NCSSs

International cooperation was discussed in all the NCSS, apart from two countries. Forty-nine countries only mentioned international cooperation as general as possible, with the rest being a bit more specific and presenting a mixture of mentions ranging from 'regional', 'bilateral' and 'multilateral' to specific countries, country blocs and/or organisations. Additionally, fifteen nations used non-binding terminology such as 'strategic (foreign) partners', 'like-minded countries', 'friendly countries', 'strategic allied countries', 'ally/allies', 'similar thinking countries', 'partners abroad', 'international partners' (Figure 6). Two outlier cases used a different approach for noncommittal terms for international cooperation; the Russian Federation, which mentioned 'interested parties' and Brazil, declaring 'the largest possible number of countries'. The most mentioned country bloc was the EU, with appearances in twenty-six individual NCSS, and the most mentioned organizations were the NATO, the UN and the OSCE, with eighteen, sixteen and thirteen mentions respectively.

International Cooperation and terms mentioned	Countries
"International" only	ARG - AUS - BGD - BLZ - BRA - CPV - CAN - CHL - CHN - COL - CRI - DOM - ECU - SLV - ETH - GTM - IND - IRL - ISR - JOR - KEN - KIR - LBN - MYS - MRT - MUS - MEX - MAR - MOZ - NZL - NIC - MKD - NOR - PAN - PRY - PHL - QAT - MDA - RUS - RWA - SAU - SEN - CHE - TJK - TUN - TKM - UKR - GBR - USA
"Regional"	AFG - BEN - CZE - SWZ - GMB - IRQ - JAM - KAZ - KWT - KGZ - MWI - NGA - ROU - ZAF - ZMB - UGA
"Bilateral" or "Multilateral"	BLR - BFA - HRV - MCO - MLT - TUR
Specific Countries (or Country blocs)	AUT - CYP - FIN - FRA - GEO - DEU - GRC - ITA - SGP - ZAF - THA
Specific Organizations	BEN - EGY - MCO - MNE - PAK - PER - SRB - LKA - UGA - VNM
Both specific Organizations and Countries (or Country blocs)	ALB - BLR - BEL - BGR - BFA - HRV - CZE - DNK - EST - HUN - ISL - JPN - KAZ - LVA - LTU - LUX - MLT - NLD - NGA - PRT - WSM - SVK - SVN - ESP - SWE - ZMB - TTO - VUT
"Strategic Partners", "Like-minded countries", "Friendly countries"	ALB - DNK - EGY - EST - GEO - JPN - LVA - PNG - POL - PRT - KOR - ROU - SWE - THA - VNM
No mention of cooperation at all	BHR - ARE

Figure 6 International Cooperation expressed in NCSSs

5. Discussion

Overall, we found that the fact that 113 (out of the 194) member states have publicly accessible NCSS shows higher interest and need for this or similar types of policy documents for member states. Given the evolving nature of technology and its effect on the government and citizens, having an effective cyber security strategy for the operation and protection of national assets, including its infrastructure and people, has become one of the key national strategic issues to address at all levels of society. As NCSSs serve to define a vision for national cyber security and, sometimes, the implementation policies and plans for all stakeholders concerned, with significant implications for coordination of both technical and personnel resources within a given timeframe, it is commendable that the majorities of member states have fresh NCSSs with a definite period of time. With most of the dates being very recent, we can say that the formulation of NCSS is a recent, up-to-date phenomenon.

Also, with a sheer number of publicly available NCSSs and oftentimes with their translations in major languages, we see that one of their purposes is of an international nature, that is, to communicate how states perceive and act in terms of cyber affairs, particularly to other states – 'like-minded' or not. Therefore, in addition to domestic coordination and resource mobilization purposes, the critical aspect of NCSSs and their publication is sharing of the nations' positions and views about cyberspace governance with foreign policy implications, and our analysis of the three key aspects ('war', 'warfare', 'conflict'/ 'neutrality', 'neutral'/ 'cooperation', 'collaboration' in the international context) has given valuable comparative insight as to states' positions and views about these aspects.

In terms of 'war', 'warfare', and 'conflict', it is recognizable from the analysis that we could almost say that it is a common understanding among most states that the threat in cyberspace from foreign states or non-state actors is potentially at the scale and scope of 'war', 'warfare' or 'conflict'. And from this rather collective use of 'war', 'warfare', and 'conflict' in the analysis, we understood that the inference might be made in terms of why the overwhelming number of NCSSs includes the term 'international cooperation' in varieties of contexts. One way of averting such a threat of 'war', 'warfare', and 'conflict' is through, for example, helping to eliminate 'safe heavens,' in the interconnected and transborder nature of cyberspace – filling the gap where potential loopholes exist for malicious exploitation of cyber vulnerabilities in other states by extending capacity-building assistance – and by establishing CBMs for more enhanced and organized information sharing. Thus, we understood that states see the merit of international cooperation as the mitigation of potential risks for their peace and security, both domestically and internationally, and have placed the importance, which is reflected in the number of NCSSs addressing international cooperation. Nevertheless, both hypotheses are validated given that only 13 countries have expressed their cybersecurity-related actions in case of war and most NCSSs discuss international cooperation.

In terms of the large appearance of the EU as a block and NATO as an organization for international cooperation, it is undiscernible from our analysis whether, for example, the need for these collective efforts to counter a threat of 'war', 'warfare', and 'conflict' in cyberspace is the result of more intensively felt or shared threat in the EU block or among NATO member states with a suggestion of a commonly perceived 'enemy', or whether, simply, it is because more collaborative frameworks exist in the EU or NATO context. Given that it is rather conclusive that states seek international cooperation as countermeasures to the threat of 'war', 'warfare', and 'conflict' in cyberspace, further study could be conducted, in more region or country-specific context, to excavate who is collaborating or seeking to collaborate in which context, which may assist states in refining their cybersecurity strategies.

Lastly, with a minimal number of references to neutrality, more or less confined to internet or technical/technological neutrality, and no mention of law of neutrality in the meaning of IHL in cyberspace, we can validate our first hypothesis and can only conclude that it is an area where decisive positions or views in terms of strategies are not yet prevalent among states, or states are still unwilling to express their positions to the outside world, maintaining a "*persistent policy of silence and ambiguity*" (Efrony and Shany 2018, p.588) that is rather frequent in issues pertaining to the identification of state practice or views in cyberspace. Given the complexity and uncertainty of the concept of neutrality itself in cyberspace, we understand that the overview of the fact of its absence is telling us of the difficulties ahead in terms of strategizing neutrality in cyberspace context both domestically and internationally.

6. Conclusion

This paper presented a global snapshot of the national positions on conflict, neutrality, and cooperation with regards to cybersecurity, of all countries (113) that have publicly released an NCSS. As hypothesized, very few countries (13) reference their cybersecurity-related preparations and preparedness for or during times of war (even though more than half discuss warfare in other general topics), and even less (5) delve into the topic of neutrality. With 111 nations discussing international cooperation, the other side of the hypothesis is similarly proven.

Future research might delve into individual national case studies, or more in-depth bi-/multi-lateral comparative studies of these topics within the cybersecurity context. While outside the scope of this study, during its course, we also discovered the existence of International Cybersecurity Strategies in some countries. It would be worthwhile to study those and compare them to their national counterparts, or between different countries. We also realize that the topic of war would be presented in more detail in (National) Defence Cyber Security Strategies; however great care must be undertaken to separate the national and institutional documents for such future studies.

By the time of this paper is published, some of the more 'outdated' NCSS might have been updated with newer versions. For them, and all new versions that will appear in the coming years, we wish to articulate the need to present a national position on war / conflict and neutrality, given the recent flare-ups between countries, as possible deterrence, CBMs, and communication instruments. This would also mean that it would be also necessary to make NCSSs publicly available. A few countries were not included in this study due to their NCSS not being publicly accessible nor available.

Acknowledgement

This article has been partially supported by the Doctoral School in Economics and Innovation, ASTRA "TTÜ arenguprogramm aastateks 2016-2022", Project code: 2014-2020.4.01.16-0032.

References

Beecroft Nick. (2021) *The UK's Cyber Strategy Is No Longer Just About Security* [Online]. Carnegie Endowment for International Peace, https://carnegieendowment.org/2021/12/17/uk-s-cyber-strategy-is-no-longer-just-about-security-pub-86037.

Caudle Sharon L. (2009) "National Security Strategies: Security from What, for Whom, and by What Means." *Journal of Homeland Security and Emergency Management* 6(1): Article 22.

Cordey Sean and Kohler Kevin. (2021) *The Law of Neutrality in Cyberspace*. ETH Zurich.

Cornish Paul, Lindey-French Julian and Yorke Claire. (2011) *Strategic Communications and National Strategy*. The Royal Institute of International Affairs: Chatham House.

CSIS - Center for Strategic and International Studies. (No date) *Significant Cyber Incidents* [Online]. CSIS, https://www.csis.org/programs/strategic-technologies-program/significant-cyber-incidents.

Daricili Ali B. and Özdal Barış. (2018) "Analysis of the Cyber Security Strategies of People's Republic of China." *The Journal of Security Strategies,* 14(28): 1–35.

Drew Dennis M. and Snow Donald M. (1988) *Making Strategy: An Introduction to National Security Processes and Problems*. Maxwell Air Force Base: Air University

DuMont Malia. (2019) *Elements of national security strategy* [Online]. Atlantic Council, https://www.atlanticcouncil.org/content-series/strategy-consortium/elements-of-national-security-strategy/.

Efrony Dan and Shany Yuval. (2018) "A Rule Book On The Shelf? Tallinn Manual 2.0 On Cyberoperations and subsequent state practice." *The American Journal of International Law* 112(4): 583–657

eGA. (No date) *NCSI – Methodology* [Online]. eGA, https://ncsi.ega.ee/methodology/.

European Union. (2015) *REGULATION (EU) 2015/2120 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 25 November 2015 laying down measures concerning open internet access and amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services and Regulation (EU) No 531/2012 on roaming on public mobile communications networks within the Union*.

Greenwood Christopher. 1987. "The Concept of War in Modern International Law." *International and Comparative Law Quarterly* 36(2): 283–306.

Hurel Louise M. (2020) *Brazil's First National Cybersecurity Strategy: An Analysis of its Past, Present and Future* [Online]. Internet Governance Project, https://www.internetgovernance.org/2020/04/05/brazils-first-national-cybersecurity-strategy-an-analysis-of-its-past-present-and-future/.

Kolini Farzan and Janczewski Lech. (2017) "Clustering and Topic Modelling: A New Approach for Analysis of National Cyber security Strategies." *PACIS 2017 Proceedings* 126.

Lauterpach Hersch. (1952) *Oppenheim's International Law Volume 2, Disputes, War and Neutrality*. 7th edition. Longmans

Libicki Martin C. (2018) *Conquest in Cyberspace: National Security and Information Warfare, Conquest in Cyberspace: National Security and Information Warfare*. Cambridge: Cambridge University Press.

Luiijf Eric, Besseling Kim, and Graaf Patrick de. (2013) "Nineteen national cyber security strategies." *International Journal of Critical Infrastructures* 9(1-2): 3–31.

Merriam-Webster. (No date) *Warfare Definition & Meaning* [Online]. Merriam-Webster, https://www.merriam-webster.com/dictionary/warfare.

Naylor Esther, Painter Christopher and Hakmeh Joyce. (2022) *How does capacity-building make cyberspace better?* [Online]. Chatham House, https://www.chathamhouse.org/2022/02/how-does-capacity-building-make-cyberspace-better.

Newmeyer Kevin P. (2015) "Elements of National Cybersecurity Strategy for Developing Nations." *National Cybersecurity Institute Journal* 1(3): 9–19.

OECD. (2012) *Cybersecurity Policy Making at a Turning Point: Analysing a New Generation of National Cybersecurity Strategies for the Internet Economy*. Paris: OECD

Przetacznik Jakub and Tarpova Simona. (2022) *Russia's war on Ukraine: Timeline of cyber-attacks* [Online]. EP, https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733549/EPRS_BRI(2022)733549_EN.pdf.

Reveron Derek S. (2012) *Cyberspace and National Security: Threats, Opportunities, and Power in a Virtual World*. Washington D.C.: Georgetown University Press

Sabillon Regner, Cavaller Victor and Cano Jeimy. (2016) "National Cyber Security Strategies: Global Trends in Cyberspace." *International Journal of Computer Science and Software Engineering* 5(5): 67–81.

Shackelford Scott and Kastelic Andraz. (2014) "Toward a State-Centric Cyber Peace? Analyzing the Role of National Cybersecurity Strategies in Enhancing Global Cybersecurity." *New York University Journal of Legislation and Public Policy*, 2016, Kelley School of Business Research Paper No. 15-4.

Shafqat Narmeen and Masood Ashraf. (2016) "Comparative Analysis of Various National Cyber Security Strategies." *International Journal of Computer Science and Information Security* 14(1): 129–136.

Snyder Hannah. (2019) “Literature review as a research methodology: An overview and guidelines.” *Journal of Business Research* 104: 333–339.

Štitilis Darius, Pakutinskas Paulius and Malinauskaitė Inga. (2017) “EU and NATO cybersecurity strategies and national cyber security strategies: A comparative analysis.” *Security Journal* 30, 1151–1168.

Teoh Chooi S. and Mahmood Ahmad K. (2017) “National cyber security strategies for digital economy.” *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)*.

U.N. (No date) *Capacity-Building* [Online]. U.N., https://www.un.org/en/academic-impact/capacity-building.

UNIDIR. No date. *About* [Online]. UNIDIR, https://cyberpolicyportal.org/about.

Von Solms Rossouw and Van Niekerk Johan. (2013) “From information security to cyber security.” *Computers and Security* 38: 97–102.

Yannakogeorgos Panayotis A. and Lowther Adam B. (2016) *Conflict and Cooperation in Cyberspace: The Challenge to National Security*. Florida: CRC Press

Wheelen Thomas, Hunger J David, Hoffman Alan N., Bamford Charles. (2017) *Strategic Management and Business Policy: Globalization, Innovation and Sustainability*. Harlow: Pearson Education.

Wu Tim. (2003) “Network Neutrality, Broadband Discrimination.” *Journal on Telecommunication & Hightech Law* 2: 141–175.

Ziolkowski Katharina. (2013) *Confidence Building Measures for Cyberspace – Legal Implications*. Tallinn: NATO CCDCOE.

Developing Cybersecurity in an Industrial Environment by Using a Testbed Environment

Jussi Simola, Reijo Savola, Tapio Frantti, Arttu Takala and Riku Lehkonen
University of Jyväskylä, Finland

jussi.hm.simola@jyu.fi
reijo.m.savola@jyu.fi
tapio.k.frantti@jyu.fi
arttu.h.takala@jyu.fi
riku.p.lehkonen@jyu.fi

Abstract: Critical infrastructure protection requires a testing environment that allows the testing of different kinds of equipment, software, networks, and tools to develop vital functions of the critical industrial environment. Used electrical equipment must be reliable, capable and maintain a stable critical industrial ecosystem. An industrial business needs to develop cybersecurity capabilities that detect and prevent IT/ICT and OT/ICS threats in an industrial environment. The emerging trend has been to create security operations center (SOC) services to detect ICS-related threats in enterprise networks. The energy supply sector must consist of crucial elements for safe business continuity and supply chain management in the industrial sector. Threats have changed into a combination of threat types. Hybrid threats may prevent everyday industrial activities, processes, and procedures so that supply chain problems may become long-lasting and affects business continuity management.

The project CSG belongs to the (Cybersecurity governance of operational technology in the sector connected smart energy) research project consortium of Business Finland's Digital Trust Programme.

The first research paper regarding the CSG (Cyber Security Governance) project concentrates on the applied theory background of this project. The research provides a research approach for investigating cyber security at the operational and technical levels. It answers the questions of where to concentrate on OT-related cyber security research and how we aim to deploy a testbed to develop a governance model in the CSG project. The study's primary purpose is to describe the operating OT-SOC environment and analyze system requirements for optimizing situational awareness in the testbed environment.

Keywords: Cybersecurity, Operational Technology, Security Operations Center, Governance Model, Testbed

1. Introduction

Business Finland-funded CSG project will develop a cybersecurity system integrated reference model to cover the common cybersecurity solutions, processes, and architecture for operational technology environments (CSG consortium). The model will be validated in several experimental implementations. The model will enable the establishment of common and standardized capabilities towards creating a competitive advantage in the global business in securing industrial automation. The University of Jyväskylä (JYU) and Turku University of Applied Sciences (TUAS) are research partners. The University of Jyväskylä coordinates the project.

CSG (Cybersecurity governance of operational technology in sector-connected smart energy networks) Significantly increases the effectiveness and efficiency of cybersecurity of smart energy networks and other operational technology. The project aims at considerable cost savings and scalability through enhanced incident management, data gathering, and Artificial Intelligence based automation. The rapid escalation of cybersecurity threats has given rise to operations centers dedicated to handling them. The SOCs – Security Operations Centers are generally considered to be best provided as a service, especially for those wishing to concentrate on their actual business rather than mastering the multifaceted attributes of cybersecurity. Cybersecurity of operational technology (OT), also called industrial control and automation systems (ICS), has been brought into the spotlight. Maersk, Norsk Hydro, and SolarWinds are recent examples of attacks with dramatic consequences (MITRE, 2022).

The unstable and insecure situation in Europe has caused a reaction that the Western world to prepare for a more effectively changing security environment. Protecting critical infrastructure, such as the electrical grid, against hybrid threats is more important than ever in history (Idaho National Laboratory, 2018). Past decades have been relatively peaceful, but nobody knows what may happen next year. European Commission set new kinds of regulatory requirements for the EU member states. In addition, trade between Western states and Russia has almost stopped caused by sanctions.

This paper provides a research approach for investigating cyber security at the operational level and answers the question of where to concentrate on OT environment cyber security research and how we aim to deploy a

testbed to develop a governance/reference model in the CSG project. The paper emphasizes the importation of a testing platform for the CSQ project at the beginning of the research project. After the first research paper, the research will be able to continue to identify and specify the comprehensive cyber security requirements and aspects for the development of OT reference model architecture, like threat and vulnerability investigations, risk assessment, and cyber security measures in an operational environment. The project's main goal Is to develop a governance and reference model for the industry stakeholders.

2. The research objectives and background in the CSG project

2.1 Essential cybersecurity requirements

European Union Agency for Cybersecurity (ENISA) is the agency that works under the European Union purposed to achieve a high common level of Cybersecurity across Europe. ENISA contributes to EU cyber policy, aims to enhance the trustworthiness of ICT products, services, and processes with cybersecurity certification schemes, and cooperates intensively with the Member States (ENISA, 2023).

For example, Directive (EU) 2022/2555 (known as NIS2) entered into force, replacing Directive (EU) 2016/1148. The directive aims to improve the existing cyber security status across the EU in different ways by:

- Creating the cyber crisis management structure (CyCLONe). The European Cyber Crises Liaison Organisation Network, EU-CyCLONe, will support the coordinated management of large-scale cybersecurity incidents.
- Increasing the level of harmonization regarding security requirements and reporting obligations
- Encouraging Members States to introduce new areas of interest such as supply chain, vulnerability management, core internet, and cyber hygiene in their national cybersecurity strategies.
- Bringing novel ideas such as peer reviews for enhancing collaboration and knowledge sharing amongst the Member States
- Covering a larger share of the economy and society by including more sectors (more entities are obliged to take measures to increase their level of cybersecurity) (EC, 2022).

2.2 Definition of IT and OT environment in the research

Information technology (IT) and operational technology (OT) have their stories. Typically, OT is used in production environments consisting of hardware and software systems that monitor and control physical equipment and processes. The stakeholders need to know more about how operational technology-related cybersecurity should be managed from a connectivity viewpoint in their system.

There is a growing need to increase the effectiveness, efficiency, and scalability of cybersecurity of OT, leading to the adaptation of capabilities of IT cybersecurity and integration of IT and OT.

Automation and using Artificial Intelligence and Machine Learning (AI/ML) will bring a competitive advantage in answering, especially scalability. In energy solutions, sector integration means integrating various energy sectors into electricity transfer networks that increase the overall complexity of the electricity networks. Still, it also enables them to balance out each other's peaks in consumption and generation, with benefits toward carbon-neutral and flexible energy systems. Compared with legacy power systems, the smart energy networks (SEN) are envisioned to fully integrate energy distribution and high-speed and two-way communication technologies into real estate, industrial factories, smart buildings, households, and millions of power equipment to establish a dynamic and interactive infrastructure with new energy management capabilities, such as demand response. It is expected that CSG will enhance the partners' capabilities and develop common capabilities, making significant steps forward in increasing competitive advantage.

2.3 Central terms

2.3.1 C2 and Service Operational Center (SOC)

Background about the term Command and Control center often refers to operative control processes and procedures of military actions. The Security Operations Center often operates 24/7, and key functions may consist of monitoring and managing customer´s security entity, developing and implementing procedures, responding to security events, analyzing customers' network traffic, providing threat intelligence reports, and implementing other security plans targets (Vielberth et al., 2020).

2.3.2 Critical Infrastructure (CI)

The systems and assets, whether physical or virtual, are so vital to society that the incapacity or destruction of such may have a debilitating impact on security, economy, public health, safety, environment, or any combination of these matters (NIST, 2018a).

2.3.3 Industrial Control Systems (ICS) and Cyber-Physical Systems (CPS)

ICS consists of supervisory control and data acquisition (SCADA) systems that control assets and distributed control systems in different places geographically. Industrial control system (ICS) means several types of control systems, including supervisory control and data acquisition (SCADA) systems, distributed control systems (DCS), and other control system configurations such as Programmable Logic Controllers (PLC) often found in the industrial sectors and critical infrastructures. An ICS consists of control components (e.g., electrical, mechanical, hydraulic, pneumatic) that act together to achieve an industrial objective (e.g., manufacturing, transportation of matter or energy) (Stouffer K., Falco J., Scarfone K., 2011). Many Cyber-Physical System applications can cause irreparable harm to the physical system under control and the people who depend on it. In particular, protecting our Critical Infrastructures that rely on CPS, such as electric power transmission and distribution, Industrial Control Systems (ICS), oil and natural gas systems, water and wastewater treatment plants, health-care devices, and transportation networks, play a fundamental role in our society. Their disruption can have a significant impact on individuals and nations at large. Cyber-Physical Systems are often operated under automated controls, and a sophisticated cyberattack can exploit weaknesses to its advantage (Hevner & Chatterjee, 2010).

2.3.4 IIoT

A smart grid system may consist of IT, a discrete system of electronic information resources organized for collecting, processing, maintaining, using, sharing, disseminating, or disposing of information. The Industrial Internet of Things (IIoT) collects data from connected devices (for example, smart connected devices and machines) in the field or plant. Then it processes these data using sophisticated software and networking tools. The entire IIoT requires a collection of hardware, software, communications, and networking technologies (Electronical Technologies, 2016)

2.4 The testbed environment of the project

CSG project aims to advance considerably in raising the cybersecurity effectiveness, efficiency, scalability, and automation of OT environments, developing novel AI/ML techniques, proactive security assurance, data gathering and filtering, and a decision support approach.

A reference model will be developed to address cybersecurity system integration, consisting of a secure smart energy sector integration governance model and an OT governance model. The goal of this model and the above-mentioned novel cybersecurity approaches together is to form a comprehensive and pioneering ecosystem for OT cybersecurity with significant business potential in international industrial automation and smart energy markets. Use cases applied in an authentic environment make it possible to test solutions and create best practices for the stakeholders. Crucial cybersecurity elements in the project consist of a) Proactive security assurance and cognitive support. b) Data gathering and decision support automation. c) Reference model consisting of governance model (including sector integration and OT cybersecurity).

This paper provides a research approach for investigating cyber security at the operational and technical levels. It answers the question of 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. The study's primary purpose is to describe the operating OT-SOC environment and analyze system requirements for optimizing situational awareness in the testbed environment.

The research approach consists of two pillars. Firstly, we have a testbed and platform to test electrical equipment such as relays and software representing critical network elements in an energy and electricity environment. The equipment we will get from stakeholders from the industry and IT sector. We have a network traffic “manipulator” that allows us to manipulate the quality of the traffic and make it possible to make delays on those. The tested equipment is crucial for stable business continuity and enables functionalities of workable industrial electric networks and ICT networks. The software that we use comprises two main categories. Open-Source-based OT-SOC software will monitor the operations and transfer relevant information to the OT-SOC of the University of Jyväskylä. Another software class belongs to industrial equipment such as IPS devices.

The built OT-SOC (Operational Technology-Security Operational Center) environment will reflect the real-life SOC environment. Testbed -work in the Jyväskylä testing laboratory will be parallel and develop the upper-level reference model. The model will be based on the results of use cases and inner and outer level-specific requirements that are considered. The Use-case inputs from testbed work in a simultaneous environment and analysis of threat assessment-based risk scenarios will create outputs for the first proposal of the governance framework.

2.4.1 Space and Configuration

There are some requirements for the designing process of creating a testbed environment.

a. The testing environment should be situated in a quiet space where different equipment combinations are possible to test, and it should have several electrical and network connection possibilities, such as sockets. Technical and electrical requirements are crucial. There must not be any extra factors causing interference. It must be a secure space where only a few selected researchers have access rights.
b. Servers, routers, tools, computers, testing components, and equipment comprise physical elements of the testbed in the SOC.
c. Monitoring software such as intrusion detection software comprises a monitoring platform for tests. In addition, the tested electrical equipment has different kinds of software and procedure to update the firmware, for example, by using a mobile.
d. Electrical networks and wired or wireless data and telecommunication network types establish the capacity to communicate, monitor traffic flow, and share information between the devices in different ways. Communication channels such as cables, Wi-Fi, LTE, and ethernet links create Information sharing entity.

The project aims to develop capabilities to monitor, detect, mitigate, and respond to OT-related cyber-physical threats and prevent and predict possible cyber-attacks. Preventative and predictive capabilities require artificial intelligence-based features. The features of the testbed-SOC concentrate on analyzing abnormal events in equipment, such as OT devices, but also abnormal traffic between them by using SOC software and IDS-tool. It also tests the features of the OT devices.

Figure 1 shows the configuration of the test environment.

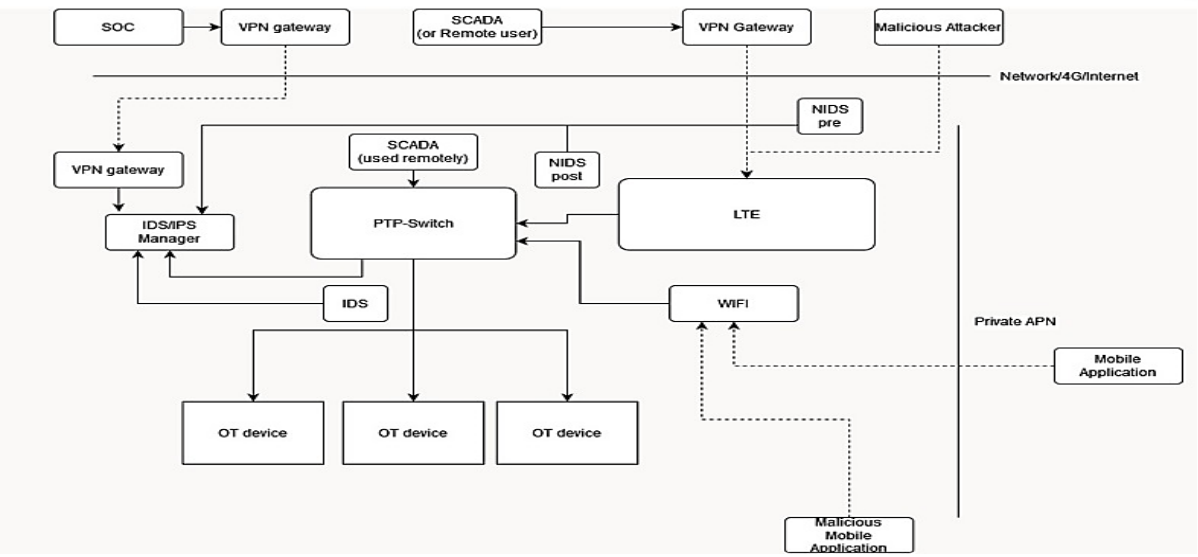


Figure 1: The testbed environment

Although the testing environment is not comprehensive, it is possible to create simulation models that demonstrate critical infrastructure-related group-based risks and scenarios that are a reality in production lines in an industrial atmosphere. It is not reasonable to test hundreds of solutions and operational technologies-based equipment combinations, but it is possible to create upper classes where to place tested use cases. The goal of the project defines the requirements for the testbed environment. There is also an “a-by project” target to produce data for the new OT-related cybersecurity standards at the industry level.

2.5 Typical OT-related Industrial threats

As Bodeau, McCollum & Fox (2018) wrote, the word threat refers to the adversary or the attack, depending on the context.

2.5.1 Industroyer

Industroyer is a sophisticated malware framework designed to causing impact the working processes of Industrial Control Systems (ICS), specifically components used in electrical substations. The industry is the first widely known malware specifically designed to target and impact operations in the electric grid (MITRE, 2022).

2.5.2 Data Destruction

Adversaries may destroy data and files on specific systems or in large numbers on a network to interrupt the availability of systems, services, and network resources. Data destruction will likely render stored data irrecoverable by forensic techniques through overwriting files or data on local and remote drives. Common operating system file deletion commands such as del and rm often only remove pointers to files without wiping the contents of the files themselves, making the files recoverable by proper forensic methodology. This behavior is distinct from Disk Content Wipe and Disk Structure Wipe because individual files are destroyed rather than sections of a storage disk or the disk's logical structure (MITRE, 2022).

2.5.3 Manipulation of Control

Adversaries may manipulate physical process control within the industrial environment. Control manipulation methods can include changes to set point values, tags, or other parameters. Adversaries may manipulate control systems devices or possibly leverage their own to communicate with and command physical control processes. The duration of manipulation may be temporary or longer sustained, depending on operator detection (MITRE, 2022).

3. Research methodological approach

The aim is to utilize several cybersecurity frameworks to define the IT and OT ecosystem in the industrial environment.

NIST framework for the Critical Infrastructure (NIST, 2018a) is a crucial general tool for the initial research. The five simple core functions of the framework are as follows a) Identify, b) Protect, c) Detect, d) Respond, and e) Recover (NIST 2018a; NIST 2022). SP 800-37 rev2. The risk management framework for information systems and organizations with the cybersecurity framework of NIST (2018b) create more detailed issues for overall risk management. NIST Special Publication 800-150 (2016) guides how to share cyber threat information.

We have used for the initial analyses cybersecurity tools of MITRE ATTA&CK (2022a, 2022b, and 2022c), 800-53rev.5 (Security and Privacy Controls for Information Systems) (NIST, 2020) and Special Publication 800-1612rev.1) Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations NIST (2022). Publication 1500-201 from NIST in Volume 2 provides an overview of the timing awareness in system elements and latency in CPS and describes special security issues concerning timing" (NIST, 2017).

MITRE's ATT&CK describe the actual attackers and their actions behind archetypes. The database provides an updatable list of detected techniques, tools, and groups. (Strom et al., 2018). ICS-related threats, actions, and tactics are also described. The database helps to form potential scenarios and consequences.

The Delphi method is suitable for conducting a relevant testbed environment analysis. The members that have been involved in this analysis process are researchers and also research methods. "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).

According to Yin (2014), the case study illustrates the attempt to produce detailed information about the object being researched. The materials collected for this case study are based on technical publications, official reports, and scientific literature.

4. Results

4.1 Where to focus cybersecurity assessment and development?

How the cyber security framework/frameworks will be used in this research project is based on an agreement with the project collaborators. The framework core provides a set of activities to achieve specific cybersecurity outcomes and references examples of guidance to achieve those outcomes. It must present key cybersecurity outcomes identified by stakeholders as helpful in managing cybersecurity risk (NIST, 2022).

We have used MITRE ATTA@CK frameworks and matrix for the threat and risk analysis work in the ICS environment (MITRE, 2020a, 2020b, 2020c). Table 1 illustrates the adversary's tactical goal in the ICS environment. Even though IT and OT environments are linked more often to each other's, they differ fundamentally from adversaries' views. Reputation and increased costs affect business continuity for all enterprises, small or big (Idaho National Laboratory, 2018). Affected OT systems are not so easy to detect. The initial research indicates that crucial vulnerabilities are related to human-based activities.

Table 1: Adversaries' tactics (Mitre 2022b)

ID	Name	Description
TA0108	Initial Access	The adversary is trying to get into the ICS environment.
TA0104	Execution	The adversary tries to run code or unauthorizedly manipulate system functions, parameters, and data.
TA0110	Persistence	The adversary is trying to maintain their foothold in the ICS environment.
TA0111	Privilege Escalation	The adversary is trying to gain higher-level permissions.
TA0103	Evasion	The adversary is trying to avoid security defenses.
TA0102	Discovery	The adversary is locating information to assess and identify their targets in the ICS environment.
TA0109	Lateral Movement	The adversary is trying to move through the ICS environment.
TA0100	Collection	The adversary tries to gather interest and domain knowledge data on the ICS environment to inform their goal.
TA0101	Command and Control	The adversary tries to communicate with and control compromised systems, controllers, and platforms with access to the ICS environment.
TA0107	Inhibit Response Function	The adversary tries to prevent safety, protection, quality assurance, and operator intervention functions from responding to a failure, hazard, or unsafe state.
TA0106	Impair Process Control	The adversary tries to manipulate, disable, or damage physical control processes.
TA105	Impact	The adversary tries to manipulate, interrupt, or destroy the ICS systems, data, and their surrounding environment.

Table 2. below indicates threat events that may affect in several ways to the OT-environment. From the view of the supply chain, the higher-level threats are more complex to govern than lower-level threats, and the consequences affect expansively.

Table 2: Examples of initial risk scenarios

Manmade events or Natural disasters	Potential consequences	Governance model scenario for Security operations (People, process, technology)
Power surges, frequency fluctuations,	Unstable operations, loss of situational awareness, production interruption, loss of reputation, cost increase	Information about frequency variation and effects, detection, mitigation, prevention, protection, repair, and information sharing with stakeholders
Electrical voltage fluctuations	Unstable operations, loss of situational awareness, production interruption, loss of reputation, cost increase	Voltage information management – detection, mitigation, prevention, protection, repair, and information sharing with stakeholders.
Structural or hostile events (physical sabotage)	Unstable operations, production interruption, loss of reputation, cost increase	Information about power supply problems, detection, mitigation, prevention, protection, repair, and information sharing with stakeholders
(cyber-attacks, man-in-the-middle, jamming, cyber-physical attack)	Unstable operations, production interruption, loss of reputation, cost increase	Information about abnormal cyber-physical events, detection, mitigation, prevention, protection, repair, and information sharing with stakeholders.
Configuration manipulation	Unstable operations, loss of situational awareness, production interruption	Information about abnormal events, detection, mitigation, prevention, protection, and information sharing with stakeholders.

Manmade events or Natural disasters	Potential consequences	Governance model scenario for Security operations (People, process, technology)
Third-party software or equipment.	Information sharing for unknown actors-unstable operation, loss of situational awareness, loss of reputation, loss of service reliability	Information about abnormal events, detection, mitigation, prevention, protection, and information sharing with stakeholders.
Environment (flooding, lightning, storm, earthquake)	interrupted production processes, and continuity management problems. two-way effects	Information about abnormal cyber-physical events, detection, mitigation, prevention, protection, and information sharing with stakeholders.
Remote update	parameters changes, unstable operations, loss of situational awareness, interrupted production processes	Information about abnormal events, detection, mitigation, prevention, protection, and information sharing with stakeholders.

The NIST (2015) specifies relevant key factors and requirements that drive design decisions regarding the control, communication, reliability, and redundancy properties at the ICS system level as follows: Control timing requirements, geographical distribution, the hierarchy of supervisory control, control complexity, the system's availability (reliability) requirements, the impact of a failure of control, the system's safety requirements.

4.2 Risk Management and Governance

There are many possibilities to understand the term governance or governance model. The primary function of a governance model is to organize the strategic, operational, and tactical/ technical level cybersecurity functionalities based on regulations, policies, guidelines, standards, and protocols. The cybersecurity governance model must consider the workplace culture, financial factors, risk and threat management, other security plans, and reporting processes that already exist in a company so that the board receives all the information it needs to put the goal of governance into practice.

The cybersecurity governance model must answer the question, how does an organization control and manages its security? Defining the risk-taking level, establishing responsibility frameworks, and determining responsibility for decision-making are essential parts of designing a governance model. The initial research indicates that the designing process of the governance model cannot be a separate part of the testbed results. There should be limitations, parameters, procedures, processes, how to deal with the different kinds of information-sharing situations, and how to share the data in a form by using a procedure that everyone understands and can work by using the same methods and procedures, and processes as other stakeholders (NIST, 2016).

There should also be an information-sharing connection between security operations centers and domain-based sectors. Information-sharing entities where to share and obtain information create added value for cybersecurity risk management. If the information-sharing method and process for the public safety authorities are not established, the threat information does not flow as required by the national and EU-level regulations. The national cybersecurity authority is the relevant actor in this case.

The research indicates that every industrial component and equipment combination is not impossible to simulate. Stakeholders have hundreds of different kinds of operational lines in their plants. That does not mean the tested combination of components and equipment does not reflect the crucial problems of the industrial environment. Creating higher-level classes and types of OT-related vulnerabilities, threats, and risks is possible. In addition, separate business units located in different locations must be standardized with each other within the business sector. The research indicates that IT- and OT-SOC -functions should be combined in the same security operations centers functionalities. Keeping those threat detection capabilities under different supervision operations makes no sense. It is meaningful to use the term SOC which consists of controlling, monitoring, detecting, possible counterattacking, and mitigation features. There is a fundamental need to reach initial threat signals from the OT factory environment that may inform SOC personnel about the potential risks that are not recognized. If the enterprise uses an operational control or supervision room in the factory, there has to have a 24/7 monitored information-sharing connection to the SOC. The SOC may also work as a filter that informs and changes the entity where the component and equipment are. Predictive features slow down operations before any serious harm happens.

4.3 Industrial sector operations and cyber security governance

People, processes, and technology are the key capabilities in the OT environment, but at the same time, they include vulnerabilities. All these parts should be covered to have high-level situational awareness of relevant threats, vulnerabilities, risks, and cyber security measures. ISO/IEC 27001:2013 (ISO27001) is the international standard for information security that specifies an information security management system (ISMS). The ISMS

assists organizations in managing their information security by addressing the people, processes, and technology (International Organization for Standardization, (ISO, 2013). NIST produces a voluntary set of guidelines for cybersecurity risk management. NIST Cybersecurity Framework (CSF) supports all other essential guidelines and standards (NIST, 2018a; NIST, 2022).

4.3.1 People: Stakeholders

A new cyber security model for governance driven by knowledge of vulnerabilities, threats, assets, potential attack impacts, and the motives and targets of potential adversaries is required. Securing the cyber aspects of an interconnected system of industrial environment hosted by multiple stakeholders requires a system-of-systems view in cyber security to analyze the whole operating environment. The organizations' stakeholders should have strategic, operational, and tactical/technical views of the cyber structure of their operational environment (Simola & Pöyhönen 2022).

At the strategic level, choices are primarily related to the organization's social responsibility, reputation, and business continuity (Finnish Standards Association SFS, 2016). The measures at the operational level promote the strategic goals, and thus, comprehensive measures will increase situational awareness for holistic cyber security management. It must be based on risk assessment and analyses of the measures based on the assessment. Maintaining cyber situational awareness concerning the organization's processes makes it possible to monitor and react efficiently to risks that constitute a threat within the organization's operating environment (Simola & Pöyhönen, 2022).

Consistent and predictable results are achieved more efficiently when operations are handled and managed as interrelated processes that function as a coherent system (Finnish Standards Association SFS, 2016). Cyber security threats and risks set special requirements for these processes and the need for situation awareness at this level in addition to other operational requirements. The targets can be achieved by defining the processes to be protected, choosing process control mechanisms successfully, and by using expedient technological solutions and services to protect the processes in the cyber environment (Stouffer et al., 2011).

4.3.2 Processes of industrial operations

Processes are key to the implementation of an effective cybersecurity strategy. Processes are crucial in defining how the organization's activities, such as governance, roles, and documentation, are used to mitigate the risks to the organization's information. Processes also need to be continually reviewed: cyber threats change quickly, and processes must adapt. People have to follow them correctly (Dutton, 2017).

4.3.3 Technology: Industrial systems

Technology is a crucial factor in the operational processes in an industrial environment. By identifying the cyber risks that an organization faces, start to look at what controls to put in place and what technologies you'll need to do this. Technology can be deployed to prevent or reduce the impact of cyber risks, depending on your risk assessment and what you deem an acceptable level of risk (Dutton, 2017).

Protecting systems, working methods, and procedures at strategic, operational, and tactical levels, security functions, and defense-in-depth strategy elements of stakeholders have to identify correctly. (Homeland Security, 2016). Effective working environments require effective cyber threat prevention mechanisms, but in addition to this, security operations must be able to eliminate combined cyber-physical threats.

5. Discussion and Conclusions

Firstly, this paper presents a research framework for the testbed work of cyber security in industrial operations and processes. It has two pillars and several research tasks. Testing platforms concentrate on practical tests, and the abstract base of research supports constructive development work. The Workable ICT and ICS systems are crucial factors for stable business continuity and risk management of the operational processes. Stakeholders must have strategic, operational, and tactical/technical level standardized security plans and guidelines consisting of cybersecurity sections to perform their daily working routines as securely as possible. Companies' coherent decision-making requires a common understanding of situational awareness about the situation from the industrial environment. Built testbed environment makes designing a cybersecurity structure for the repeatable and transferrable reference model possible. The Applied testbed environment supports to development of holistic cyber security architecture for the OT environment by using the guidelines mentioned

earlier in this paper. The proposed testbed solution is suitable for reflecting real-life operations from a real industrial environment.

Security elements like people, processes, and technology are the key capabilities in the cyber environment, but at the same time, these capabilities also include vulnerabilities. Vulnerabilities and threats can be divided into internal and external sources. Situation information from industrial production processes, relevant stakeholders, and technologies behind equipment, like machines and logistic systems, are needed to cover risk assessment in an industrial environment. This procedure is essential for monitoring IT and OT systems and sub-systems' health. The SOC that handles cyber-physical threats from the industrial control ecosystem is needed in the industrial environment and industry sector.

Technological development will challenge traditional physical monitoring systems and information-sharing methods. In the future, more than the human capacity-based ability to follow the flow of threat information or abnormal events from simple screens are required. Automated artificial intelligence-aided constitutes an opportunity to maintain supply chain and continuity management in control.

Requirements for threat visibility require gathering and analyzing data from the industrial environment and potential threats from outside. Standardized OT environments for digitalized systems and processes, procedures, data storing, data handling, etc., are needed in the near future. Efficient strategical, operational, and tactical situational understanding require SOC functionalities covering all system enterprises. Threats and vulnerabilities have to tackle before they affect a production line at the factory. All these issues made it possible to answer the research questions: where to concentrate on OT -related cyber security research and how we aim to deploy a testbed to develop a governance model in the CSG project.

References

Bodeau, D. J. & McCollum, C. D. (2018) System-of-systems threat model. The Homeland Security Systems Engineering and Development Institute (HSSEDI) MITRE: Bedford, MA, USA

ENISA, (2023. About Enisa. https://www.enisa.europa.eu/about-enisa

European Commission (2022) Directive (EU) 2022/2555. https://www.enisa.europa.eu/topics/cybersecurity-policy/nis-directive-new

Department of Defense. (2011) DoD Dictionary of Defense of Military and Associated Terms.

Dutton, J. (2017) Three pillars of cyber security. -09-26T10:00:51+00:00, [viewed April 2, 2023]. Available from: https://www.itgovernance.co.uk/blog/three-pillars-of-cyber-security.

Electrical Technology. (2016) Internet of things (IoT) and its applications in the electrical power industry http://www.electricaltechnology.org/2016/07/internet-of-things-iot-and-its-applications-in-electrical-power-industry.html.

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 25.4.2023

Finnish Standards Association SFS (2016) Johdanto laadunhallinnan ISO 9000 -standardeihin. Available on 3 April 2023: slideplayer.fi/slide/11133323/

Hevner, A., & Chatterjee, S (2010) Design research in information systems: Theory and practice. Springer Science and Business.

Homeland Security (2016) Recommended Practice: Improving Industrial Control System Cybersecurity with Defense-in-Depth Strategies Industrial Control Systems Cyber Emergency Response Team.

Idaho National Laboratory (2018) Cyber Threat and Vulnerability Analysis of the U.S. Electric Sector

International Organization for Standardization (ISO) (2013) ISO/IEC 27002:2013 Security techniques — Code of practice for information security controls. ISO. Available from: https://www.iso.org/standard/54533.html.

MITRE (2022a) "ATT&CK®for Industrial Control Systems"[online], https://attack.mitre.org/matrices/ics/

MITRE (2022b) "tactics," [online], https://attack.mitre.org/tactics/enterpise

MITRE (2022c) "ATT&CK Matrix for Enterprise," [online], https://attack.mitre.org/.

NIST (2015) SP800-82 Guide to Industrial Control Systems (ICS) Security

NIST (2016) SP 800-150 (2016) Guide to Cyber Threat Information Sharing

NIST (2017) SP 1500-201 Framework for Cyber-Physical Systems: Volume 1, Overview

NIST (2018a) Framework for Improving Critical Infrastructure Cybersecurity

NIST (2018b) SP 800-37 rev2. (The risk management framework for information systems and organizations with the cybersecurity framework)

NIST (2020) 800-53rev.5 (Security and Privacy Controls for Information Systems)

NIST (2022) NIST SP 800-1612rev.1) Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations

Simola, J., and Pöyhönen, J. (2022) Emerging Cyber Risk Challenges in Maritime Transportation. In R. P. Griffin, U. Tatarand, & B. Yankson (Eds.), ICCWS 2022: Proceedings of the 17th International Conference on Cyber Warfare and Security

(pp. 306-314). Academic Conferences International. The proceedings of the international conference on cyber warfare and security.

Strom, B.E., Applebaum, A., Miller, D.P., Nickels, K.C., Pennington, A.G. and Thomas, C.B. (2018) “Mitre Att&ck: Design and Philosophy”, Technical report.

Stouffer K., Falco J., Scarfone K. (2011) NIST Special Publication 800-82. Guide to Industrial Control Systems (ICS) Security. Recommendations of the National Institute of Standards and Technology. U.S. Department of Commerce. Available on 3 December 2021: http://csrc.nist.gov/publications/nistpubs/800-82/SP800-82-final.pdf

Vielberth M., Böhm, F., Fichtinger I. and Pernul G. (2020) "Security Operations Center: A Systematic Study and Open Challenges," in IEEE Access, vol. 8, pp. 227756-227779, doi: 10.1109/ACCESS.2020.3045514.

Yin, R.K. (2014) Case Study Research, Design and Methods. 5th ed. Thousand Oaks: Sage Publications.

Smart Terminal System of Systems' Cyber Threat Impact Evaluation

Jussi Simola, Jouni Pöyhönen and Martti Lehto
University of Jyväskylä, Jyväskylä, Finland

Jussi.hm.simola@jyu.fi
jouni.a.poyhonen@jyu.fi
martti.j.lehto@jyu.fi

Abstract: Systems of system-level thinking is required when the purpose is to develop a coherent understanding of the ecosystem where every user and system requirements are divided into specific parts. The smarter project, as a part of the Sea4value program of DIMECC, aims to develop harbor operations, including passenger and cargo transportation, in a way that port processes will improve, emissions will decrease, and overall security will enhance in smart ports. This paper describes cyber-attack impacts against the Smart terminal system of systems in the cyber realm by utilizing the MITRE ATTACK® framework to map the objectives of threat actors. The Smart Terminal system environment includes ICT, ICS networks and components, communication systems, and port service systems. Internal and external threat sources or actors are hard to divide exactly because of the diversity of the threats. Hybrid threats challenge maritime domain awareness globally. The cyber threat impacts on IT and OT environments are connected to each other because of the use of internal and external networks that impact each other by combining vulnerabilities and threats. Well-working port and terminal operations require not only protected operational systems or sensor systems, but human errors must also be minimized. Objectives of threat actors are presented, categorized, and listed. Threat scenarios illustrate that cyber threats and risks are mainly similar in the maritime global-linked port community and basic hinterland trade. The networked supply chain of the business causes evolving and combined threat scenarios. European and international standards, regulations, policies, recommendations, and, e.g., guidelines by the IMO, set new cyber-threat requirements for port and terminal services and facilities. Therefore, overall security must be considered when cyber-security is the development area. Information exchange in an understandable form is essential for maintaining business continuity. Threat information has to be transferred among stakeholders as well as cyber security codes have to be followed in the port operations of partners that are involved, for example, in operational and system-level actions. Digitalization in smart ports and terminals enhances the capacity to handle cargo and passengers more efficiently, but cyber threats evolve.

Keywords: Business Continuity, ICT, ICS, Threat (Impacts), Cyber Ecosystem, Terminal Systems of Systems

1. Introduction

This research belongs to the cyber security research actions of the SMARTER. The project goals are conducted to the reduction of emissions by optimizing harbor operations and improving cargo and people flow while improving the experience for all stakeholders. (DIMECC, 2020).

The harbor environment or maritime domain is a more changeable and challenging cyber security environment than other domains in urban areas. The diversity of the port ecosystem creates challenges for the stakeholders to maintain cyber situational awareness as a part of the port overall situational awareness. It is not enough that necessary Information and Communication Technology (ICT) and Industrial Control Systems (ICS) or Operational Technology (OT) systems aided port facilities and functions such as communication, equipment operation, cargo, and other internal and external business work independently. Separated systems challenge cybersecurity management as a part of overall security & safety management. Well-organized port governance with a business continuity strategy plan has to be implemented in an upper-level framework for the other plans of security & safety management -sectors. For example, human resources plans must be in a form that has a connection to cybersecurity plans. Human resources are a crucial part of overall cyber security, where human errors are essential factors in the cyber-physical threat world. Individual skills and abilities to percept the environment must be considered in continuous work education. Threat and risk assessments have to be clearly defined. Port authorities are crucial actors in the area of cybersecurity, The instance that owns and govern the harbors area is responsible for overall security being realized, and the maritime ecosystem stays as safe as possible. That requires shared situational awareness of the maritime domain and supply chain dependencies.

Cyber threats have risen to a top threat list in harbor areas because of digitalization, transport volumes, foreign political change, and business transformation (Atlantic Council, 2020; ENISA, 2019). Enhancing transportation and people flow is not straightforward because we must protect all procedures and processes that new political risks and threats may cause. An unstable political atmosphere expands the need for new threat-prevention mechanisms. The protected maritime domain is the crucial entity for securing intelligent systems that use multiagent AI -solutions in ports and terminals. The cyber-physical security of ports is emphasized because of

the development of political polarization. Core stakeholders and all other operators involved should follow common guidance in the future.

This fourth Smarter paper is the next step of our research process and handles phenomena of cyber threat impacts and risk scenarios. The paper concentrates on the importance of threat impact awareness in the terminal and the port process. It gives an answer at this phase of the research by using the question of what objects of cyber-attack impacts in Smart Terminal System of Systems have to take into account in cyber threat impact evaluation. The research will specify the comprehensive cyber security aspects to architect risk scenario assessment and cyber security measures for the SMARTER project. This paper is one of the outcomes of the project's final report.

2. Central concepts related to smart ports and terminals

2.1 IAPH and ICCA

The International Association of Ports and Harbors (IAPH) was founded in 1955. Member ports of IAPH handle over 60 percent of global maritime trade and around 80 percent of world container traffic. IAPH has a consultative Non-Government Organization (NGO) -based status with several United Nations agencies (IAPH, 2022a). Authorities of the state define the Port area and the ISPS -Code defines its facilities in which maritime and other activities occur (IET, 2022). Almost equal old association to IAPH, The International Cargo Handling Coordination Association (ICHCA) is dedicated to improving the safety, security, sustainability, productivity, and efficiency of cargo handling by all modes and through all phases of national and international supply chains. ICHCA International's privileged NGO status enables it to represent its members and industry at large in front of national and international agencies and regulatory bodies, including IMO. The International Technical Panel of ICHCA also provides technical advice and publications on a wide range of practical cargo handling issues. (IAPH, 2022b).

2.2 BIMCO

It is the most prominent international organization representing the interests of ship owners, charterers, brokers, and agents. The Bimco's primary role is the preparation of global regulations and policy recommendations in many areas related to the MTS, from the environment, crew support, and insurance to maritime safety and security and digitalization, including guidelines for maritime cybersecurity. (Atlantic Council, 2021).

2.3 IMO and ENISA

The maritime agency of the United Nations, the International Maritime Organization (IMO) mission is to develop a regulatory framework for international shipping. The IMO Maritime Safety Committee released a set of Maritime Cyber Risk Management recommendations for safety-management systems that IMO recommended shippers implement before the first annual verification of a vessel's Document of Compliance and Safety Management in 2021. (Atlantic Council, 2021; IMO, 2021). The European Union Agency for Cybersecurity (ENISA) is the EU's lead agency for common standards of cyber defense throughout Europe. It has introduced four cyberattack scenarios at the port community level as table 1 demonstrated (ENISA, 2019).

Table 1. Cyber-attack Scenarios selected by ENISA (2019).

Scenario A	Compromising on critical data to steal high-value cargo or allow illegal trafficking through a targeted attack
Scenario B	Propagation of ransomware leading to a total shutdown of port operations
Scenario C	Compromise of Port Community System for manipulation or theft of data
Scenario D	Compromise of OT systems creating a major accident in port areas

2.4 Maritime domain awareness

The maritime domain consists of several maritime-based sectors that create the interacting entity. Achieving common Situational Awareness (SA) requires shared situational awareness that consists of similar unchanged elements at every stage. Separate sectors of the maritime domain cannot cooperate by forming their understanding of the atmosphere independently from other actors. Therefore, a crucial factor in undistributed continuity management is common maritime situational awareness. It is helpful to classify sector-based situational awareness for creating an understandable entity. System of system-level thinking depends on the

human ability to understand the dependencies of supply chains. If we can't make a network regarding the system and business dependencies, it is challenging to develop a framework from the smaller components.

It is essential to divide different parts of situational awareness: technological SA.– organizational SA. – SA of human resources – SA. of business management – SA. of transportation – SA of regulations and policies. If all segments are well-defined and linked to each other in a way that information sharing and exchange support core functions, shared situational awareness could be achievable. Common SA. differs from Shared SA. In a concept meaning, common means a level of understanding.

Stakeholders of the harbors have to create a preliminary risk assessment where every potential threat has been considered. Previous studies related to Simola & et al. (2021) realized hybrid threats where cyber and physical risk elements are combined based on crucial human factors. It is not appropriate that risk classifications have been done separately from other risk assessments. The cyber risk assessment is an essential part of the overall risk assessment in the port. **Figure** 1 illustrates how the critical elements of the port have to analyze precisely. The risk management framework for the port functionalities may support the decision-making process, in which essential operators of infrastructure or partners undertake to cooperate in influencing the selection of risk management measures. It can be tailored to different operating environments and applies to all threats (DHS, 2013). Risk management is the "process of identifying, analyzing, and communicating risk and accepting, avoiding, transferring, or controlling it to an acceptable level at an acceptable cost (DHS, 2013). In this study, the last three sections of the framework are under consideration for risk and continuity management. We cannot eliminate threats, but we can eliminate their realization and manage them. Analyzing potential threat scenarios and objects helps to identify the tools and solutions what to use in proactive cyber-physical threat prevention mechanisms (the green and blue arrow).

Figure 1 Critical Infrastructure Risk Management (Modified from DHS 2013)

Using the simple risk management concept allows security&safety operators (e.g., Situation Center or Security Operations Center) to focus on those threats that are likely to cause risks and to use approaches designed to prevent or mitigate the effects of these potential incidents. Cybersecurity and security plans have to be a part of overall risk management and continuity management activities, where policies, processes, and procedures are defined and implemented.

3. Port and terminal system of systems

Port systems consist of several systems that handle information sharing between ICT and ICS/OT systems. Fully automated terminals exploit artificial intelligence systems by using smart sensor technology. System complexity creates challenges because several functionalities are crucial for the maritime entity. Well-scheduled port processes are a vital element for the flowing operations. For example, If timetables are not synchronized, delays may happen between the operators. Interrupted extension connections from harbors by trains and airplanes affect the whole transport supply chain. If a potential cyber-attack risk scenario is realized, functional cargo and passenger traffic faces challenges. Delays impact all stakeholders' continuity management.

3.1 Managing cyber security & Security operation centers (SOC)

Cyber Security Assessment (CSA) & Cyber Security Plan (CSP) are essential elements when establishing a cyber security management framework as part of a business continuity management framework. According to (IET, 2016; U.S Coast Guard, 2020) operational arrangement includes, for example:

- The identification of the individual(s) responsible for the cyber security of the ports and port facilities. The responsible Cyber Security Officer (CySO) is responsible for ensuring the development and maintenance of the Cyber Security Plan (CSP) and implementing and exercising the CSP.

- Port Security Committee (PSC) is needed. It is one possible way to manage stakeholders who can aid in ensuring the security of port facilities. The scope of the committee should include cyber security. The development and implementation of security procedures and measures can be enhanced by forming a Port Security Committee (PSC) (IET, 2016; International Port Security Program, 2020).
- Security Operations Center (SOC)
- Arrangements for providing information to third parties (reducing risks of sensitive information).
- Arrangements for managing security incidents and breaches (handling security breaches).

The importance of SOC´s (name varies depending on the purposes of the centers) in harbors has risen as the potential for hybrid threats increases in the maritime domain (U.S Coast Guard, 20020). Cyber and physical threats as a part of overall security management must be understood so that security personnel and the cyber emergency response team maintain shared situational awareness based on joint guidance and codes. The SOC has a centralized role as a dealer of security issues, including cyber security aspects that affect a port and port facilities. It may form a part of operations, an operations center supervising the port, controlling access, and managing business continuity and disaster recovery. The main key functions may be a) Observing by maintaining situational awareness (understanding potential threats to port facilities), b) orientation to proactive measures, and c) decisions about actions that may be appropriate to deny further access to the port asset (IET, 2016).

3.2 System complexity of port community system (PCS)

Port authorities have to coordinate and implement new digital technology as Artificial Intelligence solutions by improving service across supply chains. Port Community System is a common system for digital trade logistics, as Figure 2 demonstrates.

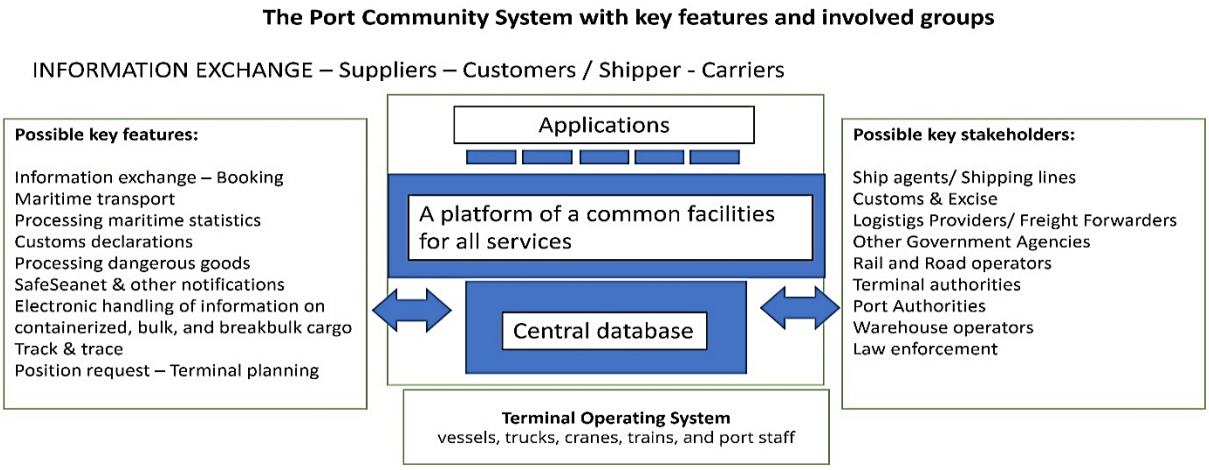


Figure 2. The Port Community System

In Finland, we do not have a long history of using PCS because our corresponding system does not have advanced digital features (IPCSA, 2022a; IPSCA2022b). We have a Portnet system by Traficom and a couple of separate systems that allow for creating situational awareness for cargo and passenger traffic in harbors. It is crucial to notice that there is globally several port -systems titled Portnet. The Port Community Systems is an open electronic platform that connects existing individual and separate systems and databases of distinct companies and organizations. The main focus of the PCS is to enable secure and intelligent operational data exchange and consolidation within the port network. Two possible seaport categories involve public and private stakeholders. a) Users who send information (e.g., shipping lines and agents, freight forwarders, and logistic actors). b) Official entities who are receiving information (port managers and operators, customs authorities, government agencies as safety authorities) (IPCSA, 2022; Sinay, 2022).

The Finnish system titled Portnet does not correspond to the requirements of international smart port system development. Finland's decentralized system governs all traffic data concerning arriving vessels, their containers, and port and terminal activities. For example, the ship reporting system Gofrep, that is developed with the neighboring county in the east, is used as a separate tool with Portnet and other operational tools that collects the vessel's information in the Gulf of Finland area (Gofrep, 2020). The Portnet has been one kind of inefficient Finnish version of the PCS. The Portnet II is under development. It should consist of more features to achieve the requirements that IMO and other international security organizations have set.

3.3 Risks, threat, and Vulnerabilities

Risk means the potential for abnormal processes or adverse circumstances of event outcomes (IAPH, 2020). Risk scenarios illustrate them. Vulnerabilities refer to the quality of state of being exposed to the possibility of being attacked or harmed, either physically meaning or emotionally meaning. **Error! Reference source not found.** above illustrates potential cyber and physical threat types in ports.

Table 1. Categorization of Emerging Threats and Vulnerabilities (George Washington University, 2021)

Threats	Vulnerabilities
Cyber	Port Infrastructure
Advanced Technologies and weapons	Automation
Violent Extremism	Port Operations
Unmanned Aerial Systems	Opportunities for Smuggling/Trafficking
	Human Factors

The threat may be an action or event that can, through the exploitation of ICT and ICS/OT, or communications infrastructure vulnerability, cause a risk to become loss or damage. Cyber-physical threats form potential risks for port infrastructure and port operations. Combined threat platforms consisting of ICT and ICS/OT and Industrial Internet of Things (IIoT)-create a potential goal for “lonely wolves” and state-level attackers.

3.4 Method of the research

This paper describes attack impacts against the Smart terminal system of systems in the cyber realm by utilizing the MITRE ATTACK framework (Mitre, 2022a; Mitre, 2022b) with official reports and other publications for analyzing the realized worldwide cyber threats and scenarios. We also used Framework for Improving Critical Infrastructure Cybersecurity, which is included in the section (analyze) in DHS's Critical Infrastructure framework. The NIST Cybersecurity Framework, and NASA's Risk-Informed Decision-Making (RIDM) framework to map and classify the objectives of threat actors and threat scenario impacts (NIST, 2018a; NASA, 2015). Concentrating on the background factors of recognized risks creates the basis for threat and risk management. The research aimed to analyze threat impacts and potential risk scenarios. The Delphi method we used is a very common cooperation model among experts and specialists. Experts use research experience in their work and assessment of the research data based on a systematic analysis of the research target.

4. Threat scenarios, consequences, and objects

Cyber-attacks set challenges not only to smart port systems but also to the whole business supply chain. For example, the scenario of Rotterdam Smart Port's intelligent features exposes it to huge vulnerabilities. Or in another scenario, APM systems which is a daughter company of maritime multi-sector operator Maersk affected by Notpetya ransomware that interrupted all terminal services in the port. The Harbor of Rotterdam has a traditional and new smart-based terminal available. The old terminal is not fully automated and can also handle containers manually. Option for another, for example, container handling method or feature (for example, manual), must be possible also in new smart ports despite the trend of digitalization. The consequences developed into a widespread chain reaction due to the form of multidisciplinary enterprise. Several ICT systems of separate business units went down, consisting of a thousand computers (George Washington University, 2021). The situation where all processes are set down causes continuity management problems and sets possibilities for added expenses and sanctions to stakeholders. For example, the customer or buyer may have the right to penalize the product sellers, transportation companies, port operators, or other stakeholders involved in the transport supply chain, especially if cyber security is mismanaged. Realized threats and their consequences often start new potential threats in new business areas. Digitalization with sensor technology requires backup & recovery systems and alternative options to use equipment in the port area.

Cyberattacks against port systems; XXX means very high supply-chain impact level, XX means high impact, and X means moderate impact, as the table illustrates. The analyzed cyber-physical risk scenarios are listed as table 3. shows. The results of the table are comparable with threat scenarios in the port and terminal system of the systems in Finland.

Table 2. Cyber-physical risk scenarios and impact rate (Mitre, 2022a).

Victim/ Place	Attacker	Type of attack	Impact/ Impact rate X-XXX	Consequences /chain-reaction	Period
Hurtigruten public transporter, Norway (Coffey H., 2020; Crew-Center,2020'; DigitalShip,2021; Naveen G.,2022; Stormshield, 2021).	Russian military hackers	Ransomware	Ransomware blocks access to files, internal email and websites. Phone lines went unavailable, and passengers' sensitive data were leaked. **Impact Rate XXX**	Loss of Business Continuity management, global I.T. infrastructure affected, major financial consequences. /Several sectors of vital functions fall under disturbances (e.g., parliament, telecommunication companies), customer cancellations, decrease in profitability, and customer relationships. Disrupted supply chain	Several months
Port of Antwerp, Belgium (Seatrade Maritime News, 2013; Stormshield, 2021).	Drug cartel group	Industrial Espionage by a keylogger.	Hijacked container management system. Allowed hackers to record the keystrokes used by the loading/unloading operators. **Impact Rate XXX**	Several containers disappeared without explanation. Hackers broke offices by deploying computers. /The supply chain and comprehensive information security were compromised.	Two years, starting from data phishing
Port of Los Angeles (BBC,2022; Greenberg, A. 2020; CBS,2017; Stormshield, 2021).	Russian military hackers - Sandworm Team/APT29	Modified (Petya) Notpetya ransomware	It affected multiple sites and selected business units. **Impact Rate XXX**	Operations of APM Terminal owned by Moller-Maersk halted. /Cargo moving operations around the docks on the landside stopped	Weeks
Rotterdam – Holland - Smart port/ APM terminal division (17 container terminals) (Dutch News, 2017; Greenberg, A. 2020; Stormshield, 2021).	Russian military hackers - Sandworm Team/APT29	Modified (Petya) NotPetya ransomware called Petrwrap	Ransomware blocks access to computer-based systems and cranes. Destroyed computer systems. **Impact Rate XXX**	The fully automated terminal went disabled. Loss of Terminal services disturbed Cargo handling and transport chain. /The operational functions of several companies were interrupted. E.g., containers and daughter companies of Maersk. Disrupted supply chain.	Several weeks, despite the backup and recovery methods.
Cosco operations in Port of Long Beach, USA. (Seatrade Maritime News, 2018; Stormshield, 2021)	Russian military hackers - Sandworm Team/APT29	Series of international NotPetya ransomware attacks	Disrupted the activities of several international ports. **Impact Rate XX**	Access was denied to the U.S. website, stoppage on the email and phone. /Email communication problems with carriers' U.S. operations and its customers. Impact on WAN and VPN gateways.	Weeks
Port operator of Barcelona, Spain. (DHS, 2019; Stormshield, 2021).	Unknown actor	Ransomware	Affected internal land operations and its systems, disrupted e.g., loading and unloading of boats. **Impact Rate XX**	Transport operations stopped.- The potential connection between case Barcelona and case San Diego cyber attack	A few Days
Port operator of San Diego, USA. (DHS, 2019; San Diego Union Tribune, 2018; Stormshield, 2021).	Iranian Cybergang	Samsam Ransomware - a highly sophisticated cyberattack	Public agency's ability to process and perform services discontinued. **Impact Rate XX**	The attack shut down port services-ability to pay traffic tickets and bills./The port has an integral role in public safety via connection to the Harbor police. The same attacker disrupted wireless communications at Atlanta airport.	Weeks
Vancouver, Canada. (Pesanti, D., 2017; Stormshield, 2021)	Unknown actor	A Distributed Denial of service attack (DDoS) - Brute force attack	Affects sending many work requests in the systems. **Impact Rate XX**	225000 user accounts were propped./Wifi-connected computers caused the spread of the virus. Networked and Connected computers were affected.	Months
Marseilles, France (CERT-FR,2020; MITRE;2022a.	Opportunistic	Ransomware Mespinoza/Pysa	PYSA ransomware is a ransomware-as-a-service (RaaS) tool that disables	Interconnected information systems with Aix-Marseille-Provence/ Organizations in	A couple

Victim/ Place	Attacker	Type of attack	Impact/ Impact rate X-XXX	Consequences /chain-reaction	Period
; Stormshield, 2021)	unknown actor		some security solutions. **Impact Rate XX**	Provence caused a chain reaction within vital functions.	of weeks
Vard of Fincantieri, Langsten, Norway. (Stormshield, 2021)	Unknown Actor	Ransomware attack	**Impact Rate unknown**	Has declined to give details.	-
Kennewick, USA. (Maritime-Executive 2020; (Stormshield, 2021;Wingrove, M.,2020).	Military hackers	Ransomware demanded $200,000 in ransom to restore access to the port's servers and files.	Criminals locked the port administration, bypassing firewalls and antivirus software. **Impact Rate XX**	Port of Kennewick was unable to use these locked servers.	A couple of Weeks
Transnet National Port Authority, South Africa. (Reuters,2021; Stormshield, 2021)	Unknown Actor	a case of cyber-force majure by ransomware	Force majeure against container terminals. **Impact Rate XXX**	Four major ports were paralyzed. /Backlogs and hamper exports from the region.	Several weeks
Port of Houston, USA. (Lyngaas, S., 2021; Donelly, J.,202); Infosecurity,2021; Stormshield,2021)	State-sponsored actor	Hackers exploit a vulnerability in password management software titled "ManageEngine ADSelfService Plus," which is used for password management and single sign-on (CVE-2021-40539).	Attackers broke into one of the port's web servers and installed malicious code to expand their access to the system. Then exfiltrated all the log-in credentials for a piece of Microsoft password management software used to control network access. **Impact Rate X**	Potential consequences if compromise had not been detected: Unrestricted remote access to the (IT.) network./Compromised supply chain operations.	Several days

Five risk scenarios achieved supply-chain impact level 3x, Table 3. Achieving this level requires widespread consequences and unexpected supply-chain impacts abroad.

Table 3. Supply-chain impacts

Norway, Case Smart Port in Hurtigruten - Ransomware	xxx
Belgium, Industrial Espionage port of Antwerpen - Supply chain attack	xxx
Holland, Rotterdam, collateral damage – NotPetya ransomware	xxx
South Africa, a case of cyber - force majeure by ransomware	xxx
USA, Port of Los Angeles – Notpetya ransomware	xxx

As research outcomes, potential threat scenarios may start from ransomware or phishing attack. Some attackers exploited weaknesses in the system, and human errors and activities have caused others. We have investigated and analyzed the most significant cyber-threat cases that are spread widely and cause major problems to the business continuity management of the enterprises. Petya-ransomware cyber-attack affected infected terminal systems so that several vessels were diverted to other terminals to ensure that customers' cargoes were not unduly delayed (MSC, 2017). Attacks against maritime supply chains have raised the most popular target by attackers. Therefore, e.g., the port of Los Angeles in the U.S. cooperates with the cyber resilience center of the FBI (BBC, 2021; Safety4sea, 2022b; Safety4sea, 2022c). Cybersecurity Supply Chain Risk Management (C-SCRM) is a usable process for managing exposure to cybersecurity risks throughout the supply chain and developing appropriate response strategies, policies, processes, and procedures. It helps the enterprise manage to cybersecurity risks throughout the supply chain (NIST, 2021). Cybersecurity supply chain risk management (C-SCRM) activities described in this publication are closely related to the Risk Management Framework described in NIST SP 800-37, Rev. 2. (NIST,2018b), SP 800-30, Revision 1, Guide for Conducting Risk Assessment (NIST, 2012). Figure 3. demonstrates how business continuity and port requirements are connected to each other. Supply chain risks are possible to tackle only by creating a common situational understanding of stakeholders operational working culture and working process/procedures.

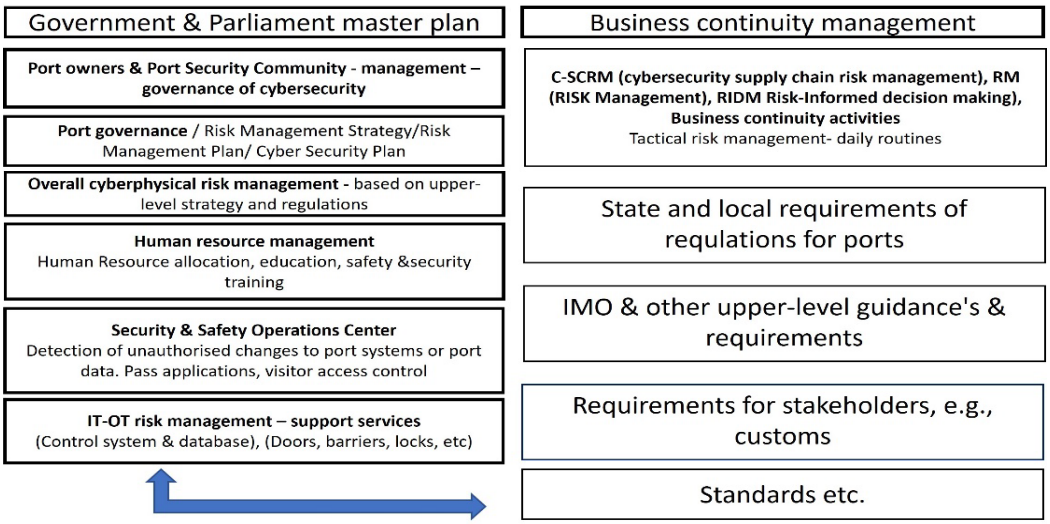


Figure 3. Supply Chain Risk Management

As research data analysis indicates, there is no cyber-threat-free or protected business sectors. Many infected companies of realized threats in terminal areas are seamlessly connected to other industries that are linked with each other through shared systems. This kind of information-sharing and communication cycle creates more possibilities for cyber attackers.

The most challenging threats to detect and potential risk scenarios in the future are related to 3rd party services NSC.gov.uk. (2019), which may cause uncontrollable supply chain attacks that evolve into the other business domain as listed above. The potential reputational harm is not limited to the company under cyber-attack; cyber-attacks against the supply chain may cause reputational damage to several business sectors as Figure 4. clarify.

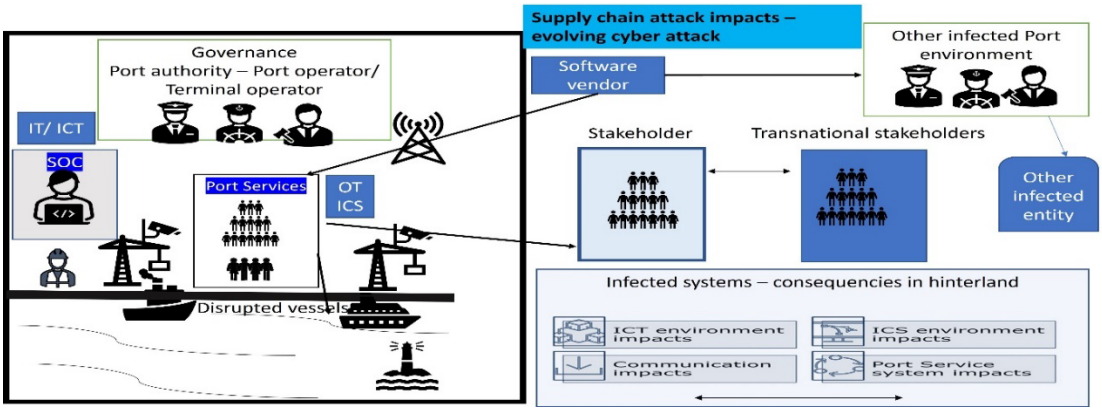


Figure 4 Supply-chain attack impacts.

The third relevant risk scenario is based on the lack of human resource management, which may be due to a lack of data protection and information security training. Lack of understanding of cyber and physical threats.

- 3d party Compromised legitimate software (hijacked ICS software)
- 3d party hardware, port equipment, cameras, drones, routers, sensors, devices, and other unknown adverse components
- Lack of human resource management - Intentional and unintentional human errors caused by lack of training, changing personnel, management of rights of access and use

5. Conclusion

As the survey indicates, it is not enough to have a Cyber Incident Response plan in port and terminal processes. To have overall protection against cyber security incidents or hybrid threats, cyber incident response plans and other security plans have to be flexible, and they have to be linked to each other and continuously updated. It means nothing to have only a plan without implementing and systematically auditing it. It seems that the same (selected) assessment tools or frameworks have to be used by all who are involved. Practical operational

fieldworkers must understand the meaning of cybersecurity and the effecting elements of it in their daily routine and working procedures. Threat information exchange via the security operation center with the nation's official cyber incident response team requires workable connection and information-sharing methods among private-public-private information sharing. Multinational and multisectoral companies have to arrange their daily operative working procedures by implementing and following guidelines for overall security management that consist of human resource & control management but also cybersecurity management of ICT/ICS/OT resources. System-level thinking requires an understanding of maritime diversity with supply chain dimensions because supply chain attacks are a crucial risk in the port ecosystem. Risk management is a crucial part of business continuity management. Aware that personnel training for maintaining cyber security is equally essential as updated software and hardware in all systems. Maintaining situational awareness in terminals and ports comprises elements that are introduced in this research. Mapping of threat scenarios aids managers and authorities in preparing against the most adverse threats. Therefore, supply chain has to see the broader framework where critical infrastructure and its vital functions are protected in a way that smart solutions such as sensor systems may support port services and facilities continuously without essential breaks in cargo transportation and passenger traffic. Port authorities, with other selected crucial actors, have to coordinate cyber security platforms in their community in a way that each level and each corporate t communicate with the same “language”. A mental model of the terms has to be at the same level. If this fundamental factor is not recognized, everything else is pointless and energy will flow to solving information-sharing problems and lack of understanding. The system of system-level thinking in the global-linked port community has to have common terms and language that form cybersecurity requirements for all stakeholders involved. Everything can be built on top of this platform, from management to responsibilities, standards, guidelines, policies, and other rules.

The paper provides a research approach to realized risks and the potential threat and risk scenarios/impacts on the port systems and facilities of those. The research approach uses the system of systems (SoS) thinking. The findings of the study propose the main cyber-physical risks and impacts of risk factors that affect business continuity management, supply chain, transport, and the whole maritime domain. The proposal indicates the issues that have to recognize in a part of the entire management system. Companies in the port area and collaborated organizations must maintain and update a risk threat scenario tool similar to Table 3. in their risk scenario and impact assessment work, which has been used to evaluate the impacts and consequences of cyber threats in the port area. It will help to focus on the potential threat impacts, threat objects, and proactive measures.

References

Atlantic Council. (2020). Raising the Colors: Signaling for Cooperation on Maritime Cybersecurity

BBC. (2022). Cyber-attacks on the port of Los Angeles have doubled since the pandemic. Https://www.bbc.com/news/business-62260272.

CBS. (2017). L.A.Port terminal still shut down following cyber-attack. https://www.cbsnews.com/losangeles/news/la-port-terminal-still-closed/

CISA. Port Facility Cybersecurity risks https://www.cisa.gov/sites/default/files/publications/port-facility-cybersecurity-risks-infographic_508.pdf

Coffey, H. (2020). Cruise company hit by “comprehensive” cyberattack. https://www.independent.co.uk/travel/news-and-advice/hurtigruten-cruise-company-ransomware-cyber-attack-b1774890.html

Crew-center. (2020). Cyberattacks on the rise: Norwegian Cruise Company latest victim on Ransomware https://crew-center.com/cyberattacks-rise-norwegian-cruise-company-latest-victim-ransomware

Customs City. (2022). What is a port community system (PCS)?. https://customscity.com/what-is-a-port-community-system-pcs/

Cybermaretique. (2020). Le croisiériste norvégien Hurtigruten victime d’une attaque par rançongiciel https://cybermaretique.fr

DCSA. (2020). DCSA Implementation Guide for Cyber Security on Vessels v1.0. https://dcsa.org/wp-content/uploads/2020/03/DCSA-Implementation-Guideline-for-BIMCO-Compliant-Cyber-Security-on-Vessels-v1.0.pdf

DHS. (2019). Northern California area maritime security committee cyber security newsletter. DHS. https://homeport.uscg.mil/Lists/Content/Attachments/40226/1901.pdf

DHS. (2013). NIPP 2013 - Partnering for Critical Infrastructure Security and Resilience.

Digital Ship. (2021). Passenger data leaked in the Hurtigruten cyber-attack. https://thedigitalship.com/news/maritime-satellite-communications/item/7147-passenger-data-leaked-in-hurtigruten-cyber-attack

DIMECC Oy, (2020). DIMECC Sea4Value/Smart Terminals (SMARTER). Project proposal for One Sea – autonomous maritime ecosystem.

Donelly, J. (2021). Port Houston targeted by suspected nation-state actor in cyber-attack. Port Technology International. https://www.porttechnology.org/news/port-houston-targeted-by-nation-state-actor-in-cyber-attack/

Dutch News. (2017). Smart port in Rotterdam confounded by cyber attack. Http://www.dutchnews .nl/news/2017/06/smart-port-in-rotterdam-confounded-by-cyber-attack-

ENISA. (2019). Port Cybersecurity – Good practices for cybersecurity in the maritime sector.

European Parliament. Directive (E.U.) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union

Fintraffic. (2020). GOFREP. https://www.fintraffic.fi/sites/default/files/2021-09/GOFREP_MG_2021_09_03.pdf

George Washington University, (2021). Emerging Risks in the Marine Transportation System (MTS), 2001-2021. NCITE

Greenberg, A. (2018). The untold story of Notpetya, the most devastating cyberattack in history. Wired. https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/

IAPH. (2020a). Port Community Cyber Security. https://sustainableworldports.org/wp-content/uploads/IAPH-Port-Community-Cyber-Security-Report-Q2-2020.pdf

IAPH. (2020b). Cybersecurity Guidelines for Ports and Port Facilities. https://sustainablewoelports.org/wp-content/uploads/IAPH-Cybersecurity-Guidelines-version-1_0.pdf.

IET. (2020). Institution od Engineering and Technology 2016. Cyber Security for Ports and Port Systems. Department of Transport.

IMO. International Maritime Organization. (2022). Maritime safety. https://imo.org/en/OurWork/Safety/Pages/default.aspx

Infosecurity. (2021). Port of Houston quells cyberattack. https://www.infosecurity-magazine.com/news/port-of-houston-quells-cyberattack/

IPCSA. International Port Community Systems Association. (2022a). Port Community Systems. Https://ipsca.international/pcs/pcs-general/

IPCSA. International Port Community Systems Association. (2022b). How to develop a Port Community System. https://ipcsa.international/publications/how-to-develop-a-port-community-system/

Lyngaas Sean. (2021). Hackers breached the computer network at a key U.S. port but did not disrupt operations.CNN.https://edition.cnn.com/2021/09/23/politics/suspected-foreign-hack- houston/index.html.

Maritime-Executive. (2020). Ransomware Cripples IT Systems of Inland Port in Washington State. https://www.maritime-executive.com/article/ransomware-attack-cripples-systems-of-inland-port-in-washington-state

MBLT. Maritime Bulk Liquids Transfer Cybersecurity Framework Profile

Mitre. (2022a). “ATT&CK Matrix for Enterprise”, [online], https://attack.mitre.org/.

Mitre. (2022b). “Attack Navigator”, [online], https://mitre-attack.github.io/attack-navigator/

Naveen Goud. (2022). Hurtigruten suffers a serious Ransomware Attack. Cyber-Security Insiders. https://www.cybersecurity-insiders.com/hurtigruten-suffers-a-serious-ransomware-attack/

NASA. (2015). Considering Risk and Resilience in Decision-Making.

NIST. (2021). NIST Special Publication NIST SP 800-161r1 Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-161r1.pdf

NIST. (2018a). Framework for Improving Critical Infrastructure Cybersecurity.

NIST. (2018b). SP 800-37, Revision 2, Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy:’’

NIST. (2017). Special Publication 1500-201 Framework for Cyber-Physical Systems: Volume 1, Overview. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-201.pdf

NIST. (2012) Guide for Conducting Risk Assessments.

NSC.gov.uk. (2019). Third-party software providers. www.nsc.gov.uk./collection/suplly/-chain-security/third-party-software-providers

Pesanti D. (2017). Port of Vancouver meeting hindered by cyberattack. The Columbian

Reuters. (2021). South Africa's Transnet restores operations at ports after cyber attack https://www.reuters.com/article/us-transnet-cyber-idUSKBN2EZ0RQ

Safety4Sea. (2022a). USCG: Be aware of typosquatting of port facility websites. https://safety4sea.com/uscg-be-aware-of-typosquatting-of-port-facility-websites-typosquatting

Safety4Sea. (2022b). Us ports and terminals targets of increased cyber security attacks https://safety4sea.com/us-ports-and-terminals-targets-of-increased-cyber-security-attacks/

Saety4Sea. (2022c). Port of LA: Cyber-attacks have doubled since the pandemic. https://safety4sea.com/port-of-la-cyber-attacks-have-doubled-since-pamdemic/

San Diego Union Tribune. (2018). https://www.sandiegouniontribune.com/business/growth-development/sd-fi-port-cyberattack-20180926-story.html

Seatrade Maritime News. (2013). Antwerp incident highlights maritime I.T. security risk. https://www.seatrade-maritime.com/europe/antwerp-incident-highlights-maritime-it-security-risk

Seatrade Maritime News. (2018). Americas coscos us operations hit by cyber attack. https://www.seatrade-maritime.com/americas/coscos-us-operations-hit-cyber-attack

Simola, J., Lehto, M., Rajamäki, J. (2021). Emergency Response Model as a part of the Smart Society

Sinay. (2020). What is port community systems? https://sinay.ai/en/what-is-port-community-system/

MSC Mediterranean Shipping Company. (2017). Customer update: Petya Cyber-Attack. Https:// www.msc.com/en/newsroom/news/2017/july/customer-update-petya-cyberattack

Stormshield. 2021. Cybermarétique: a short history of cyberattacks against ports. https://www.stormshield.com/news/cybermaretique-a-short-history-of-cyberattacks-against-ports/

U.S: Coast guard. (2020). Inter-American Committee on ports. International Port Security Program. Stakeholders management - Port security committees.

Walker, J. (2022). Ports and Terminals Cybersecurity Survey.

Wingrove, M. (2020). Riviera. Cyber attack shuts down US port servers. https://www.rivieramm.com/news-content-hub/news-content-hub/cyber-attack-shuts-down-us-port-servers-61955

A Cyber Counterintelligence Competence Framework: Developing the Job Roles

Thenjiwe Sithole, Jaco Du Toit and Sebastian H von Solms
Academy of Computer Science and Software Engineering, University of Johannesburg, South Africa

thenjiwes@icloud.com
jacodt@uj.ac.za
basievs@uj.ac.za

Abstract: In recent years, there have been intensifying cyber risks and volumes of cyber incidents prompting a significant shift in the cyber threat landscape. Both nation-state and non-state actors are increasingly resolute and innovative in their techniques and operations globally. These intensifying cyber risks and incidents suggest that cyber capability is inversely proportional to cyber risks, threats and attacks. Therefore, this confirms an emergent and critical need to adopt and invest in intelligence strategies, predominantly cyber counterintelligence (CCI), which is a multi-disciplinary and proactive measure to mitigate risks and counter cyber threats and cyber-attacks. Concurrent with the adoption of CCI is an appreciation that requisite job roles must be defined and developed. Notwithstanding the traction that CCI is gaining, we found no work on a clear categorisation for the CCI job roles in the academic or industry literature surveyed. Furthermore, from a cybersecurity perspective, it is unclear which job roles constitute the CCI field. This paper stems from and expands on the authors' prior research on developing a CCI Competence Framework. The proposed CCI Competence Framework consists of four critical elements deemed essential for CCI workforce development. In order of progression, the Framework's elements are: CCI Dimensions (passive-defensive, active-defensive, passive-offensive, active-offensive), CCI Functional Areas (detection, deterrence, deception, neutralisation), CCI Job Roles (associated with each respective Functional Area), and Tasks and Competences (allocated to each job role). Pivoting on prior research on CCI Dimensions and CCI Functional Areas, this paper advances a proposition on associated Job Roles in a manner that is both intelligible and categorised. To this end, the paper advances a five-step process that evaluates and examines Counterintelligence and Cybersecurity Job Roles and functions to derive a combination of new or existing Job Roles required for the CCI workforce/professionals. Although there are several cybersecurity frameworks for workforce development, establishing the CCI Job Roles is specifically based on the expression of the Job Roles defined in the National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework.

Keywords: cyber counterintelligence, cyber counterintelligence job roles, counterintelligence functions, cybersecurity job roles.

1. Introduction

The rapid shift in the cyber threat landscape indicates that cyber threats and cyber-attacks in various forms are a peril to technological advances and to the confidentiality, integrity and availability and availability of information systems and critical information infrastructure. Both nation-state and non-state actors are increasingly resolute and innovative in their cyber techniques and operations at a global scale. Cyber-attacks are not only increasing in sophistication, magnitude and quantity, but they are becoming more aggressive (World Economic Forum, 2022). The World Economic Forum on the Global Risks Report for 2023 flagged cyber-attacks on critical (information) infrastructures as one of the top risks possessing considerable global impact (World Economic Forum, 2023).

We are in an era where technological advancement is inevitable; it has created smart cities, the interconnectedness of critical infrastructures through the convergence of information technology and operational technology, and digital services in sectors such as government, healthcare, education, and banking. This is also an era where cyber risks, cyber-attacks and the spread of disinformation are also inevitable. Threat actors are already increasing their capabilities, using artificial intelligence-powered cyber-attacks (Guembea, et al., 2022; Jaber & Fritsch, 2023). There has been a 38% increase in cyber-attacks worldwide in 2022 compared to 2021, with education/research, government/military and healthcare being the most attacked sectors (Check Point Research, 2023). There has also been an increase in cyber disruption of critical infrastructures between 2021 and 2022, with some of the infrastructures targeted by nation-state actors (Microsoft, 2022; Trend Micro, 2022)

These intensifying cyber risks and attacks suggest that cyber capability is inversely proportional to cyber risks, threats and attacks. How, then in this era of complexity, can governments and organisations be resilient and thrive? How can governments and organisations protect the confidentiality, integrity and availability (CIA) of their information systems? It is no longer refutable that traditional cybersecurity measures are not enough. Therefore, this confirms an emergent and critical need for organisations to adopt and invest, predominantly, in

cyber counterintelligence (CCI), which is a multi-disciplinary and proactive measure that can identify, deter, prevent, destruct, exploit and neutralise adversarial attempts to collect, alter or breach the CIA of critical information systems through cyber means (Duvenage & von Solms, 2015).

Although CCI is gaining recognition, CCI is an emerging multi-disciplinary field that lacks a well-defined workforce structure and competences. It is apparent that a competent CCI workforce is needed to counter the escalating cyber-attacks and mitigate the cyber risks. However, there is no work on a clear categorisation of the CCI Job Roles or competences in the academic or industry literature reviewed. Furthermore, from a cybersecurity perspective, it is unclear which Job Roles or competences constitute the CCI field.

Based on the authors' previous research on building a CCI Competence Framework, this paper primarily aims to advance a proposition on CCI Job Roles in an intelligible and categorised manner. The Job Roles are the core of the CCI Competence Framework as they concretise the tasks and the competences required to achieve the organisational CCI agenda, as well as to be used to guide the design of education, training and development programmes.

This paper is structured as follows: Section 2 examines the concept of a Job Role and why it is essential for workforce development. Section 3 outlines and selectively illustrates the approach for identifying and creating the Job Roles for the CCI Competence Framework. This five-step process utilises, but also moves beyond, the National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework. Section 4 concludes the paper by outlining aspects for further research in refining the proposed CIC Competency framework.

2. What is a job role, and is it important?

Prior to exploring the identification of Job Roles, it is crucial to understand what a Job Role is and how it relates to workforce development because it will impact the design of the CCI Competence Framework. Job Roles refer to a comprehensive set of tasks, of which each task is associated with competences (knowledge, skills and abilities) for which someone must perform that role and their capability or competence level to perform the task is determined by the proficiency levels (Neal, et al., 2012). A Job Role has the following fundamental components:

- Job title – means the name used to refer to a particular Job Role.
- Job code – a unique code identifier used to reference the Job Role
- Job purpose – comprises an overview of what the Job Role entails: the primary purpose and objectives of the Job Role.
- Tasks – a detailed, specific list of primary responsibilities needed for the Job Role.
- Level of proficiency – description of the levels of proficiency required for each task and competences of the Job Role.
- Competences – a detailed list of all competences (knowledge, skills and abilities) required to perform the tasks.

A critical component of a well-defined job role is ensuring that organisations have the right people at every level to perform the tasks optimally. The fundamental rationale for job role identification and description is the requirement to identify, recruit, train, develop and retain an appropriately qualified and competent workforce, which is accomplished by describing the whole range of related tasks and competences (knowledge, skills and abilities) that must be performed (Saltz & Grady, 2017). Job roles will also assist education and training institutions in designing and developing relevant programs based on common taxonomy and competences (knowledge, skills and abilities) required for particular job roles in a particular industry (Hajny, et al., 2021).

This section provided an overview of what a job role entails and why a job role is important for a competence framework and workforce development. The following section explains the approach to developing the Job Roles for the CCI Competence Framework.

3. Identifying job roles for the CCI Competence Framework

In this paper, we evaluate and examine counterintelligence, CCI and cybersecurity functions to construct practical, intelligible and categorised new or existing job roles for the CCI workforce /professionals. Prior research by the authors proposed a CCI Competence Framework consisting of four critical elements deemed essential for CCI workforce development. In order of progression, the Framework's elements are CCI Dimensions (passive-defensive, active-defensive, passive-offensive, active-offensive), CCI Functional Areas (detection,

deterrence, deception, neutralisation), CCI Job Roles (associated with each respective Functional Area), and Tasks and Competences (allocated to each job role) (Sithole & Du Toit, 2022). The interlinkages between these elements, dimensions and functional areas are graphically depicted in Figure 1.

In their previous research, the authors identified and defined the first two elements of the CCI Competence Framework, namely CCI Dimensions and CCI Functional Areas (see Figure 1). As Figure 1 also shows, the construct of these two elements aligns with the concept of the four dimensions of counterintelligence (Prunckun, 2019) and the CCI (Duvenage & von Solms, 2014).

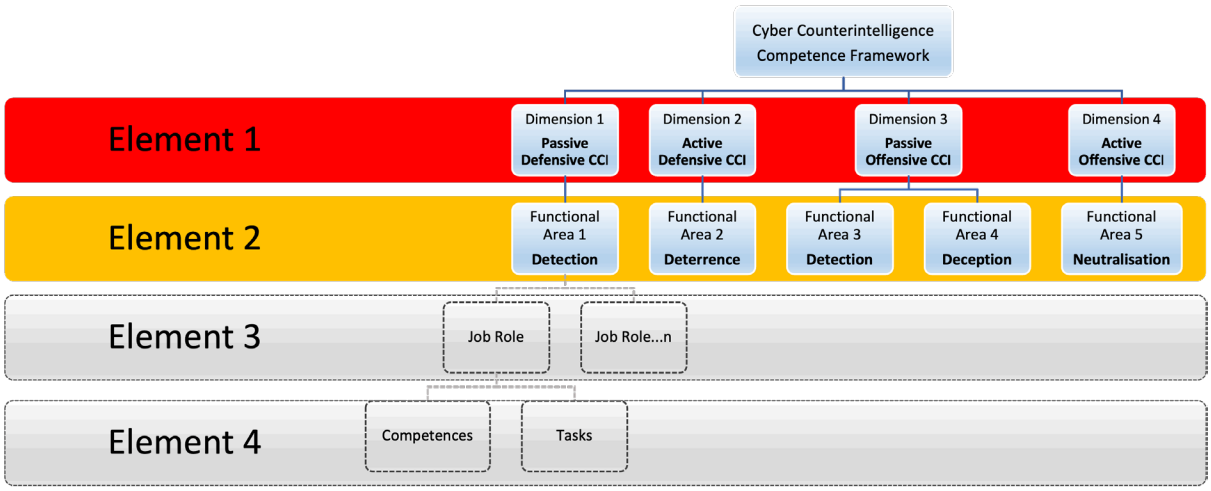


Figure 1: Elements 1 and 2 of the CCI Competence Framework - the Four Dimensions and Five Functional Areas (Source: Sithole & Du Toit, 2022)

The paragraph above demonstrated that two Elements of the CCI Competence Framework (Figure 1) had been developed in preceding research. The rest of this section proceeds with discussing the approach in building Element 3, namely Job Roles. To this end, the following sub-sections outline a five-step process to identify and create Job Roles for the CCI Competence Framework. This process commences with selecting an appropriate cybersecurity framework to use as a reference resource and 'mapping template'.

3.1 Choosing the NICE Framework

To identify the Job Roles, it is crucial to recognise cybersecurity-related skills or workforce development frameworks for their usability and relevance to achieve the objective of this paper. We identified the widely considered NICE framework, which represents the foundation of cybersecurity workforce development, and is also used as a reference source or guide for developing cybersecurity education and training programmes (Furnell, 2021; Hajny, et al., 2021; SPARTA, 2020). The NICE Framework is analysed to identify the Job Roles that are within the broad context of our CI Competence Framework and CI in general.

The NICE Framework is an initiative of the National Institute for Standards and Technology (NIST) as a foundation for cybersecurity education and workforce development. The goal of the NICE Framework is to provide a common, consistent terminology for precise categorisation and description of work roles and their associated tasks, knowledge, skills and abilities needed for cybersecurity work, education and training. (Newhouse, et al., 2017).

As illustrated in Figure 2, the NICE Framework comprises several hierarchal components – Categories, Specialty Areas, Work Roles, Tasks, and Knowledge, Skills and Abilities. The NICE Framework is arranged into seven overarching categories based on a general grouping of cybersecurity functions – Analyse, Collect and Operate, Investigate, Operate and Maintain, Oversee and Govern, Protect and Defend, and Secure Provision. Spread across these categories are 31 Specialty Areas, each having one or more Work Roles. The NICE Framework has 52 Work Roles with several associated Tasks, Knowledge, Skills and Abilities.

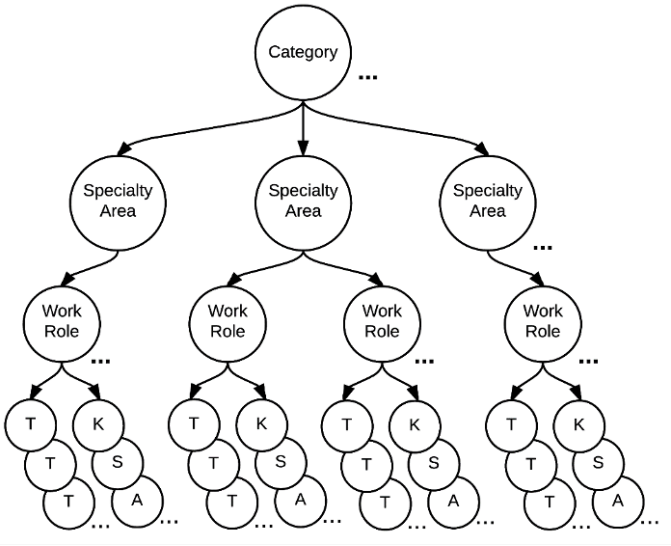


Figure 2: A hierarchical relationship of the NICE Framework components (Source: Newhouse et al., 2017)

The rationale for using the NICE Framework as a basis for identifying the Job Roles for the CCI Competence Framework is that it provides applicable, practical and detailed cybersecurity Work Roles.

3.2 Mapping of the CCI Competence Framework with the NICE Framework

CCI is a subset of CI; therefore, it is essential to note that CCI is underpinned, not only by the four CI dimensions but, also by the functions of CI – CI Analysis, CI Investigations, CI Operations and CI Collection that apply to all the dimensions of CI (Prunckun, 2019). These CI functions will also be applicable to the functional areas of the CCI Competence Framework. Considering the limited research on the CCI field in general and the blurriness of cybersecurity skills/workforce frameworks on which Job Roles and competences constitute the CCI field, the CI functions will assist when identifying the Job Roles during mapping with the NICE Framework. In accordance with its dictionary description, we deem "mapping as an operation that associates each element of a given set … with one or more elements of a second set" (Oxford Dictionary of English, 2010).

By mapping the CCI Competence Framework to the NICE Framework, we aim to identify Job Roles that are comparable to the CCI Competence Framework Functional Areas, and the functions (analysis, collection, investigation and operations) and activities (identity, deter, prevent, destruct, manipulate, exploit and neutralise) of CI and CCI. A top-down approach is found to be the most suitable for identifying or creating Job Roles for the CCI Competence Framework through the mapping process. The 5-step mapping process used is depicted in Figure 3, and a detailed description for each step follows.

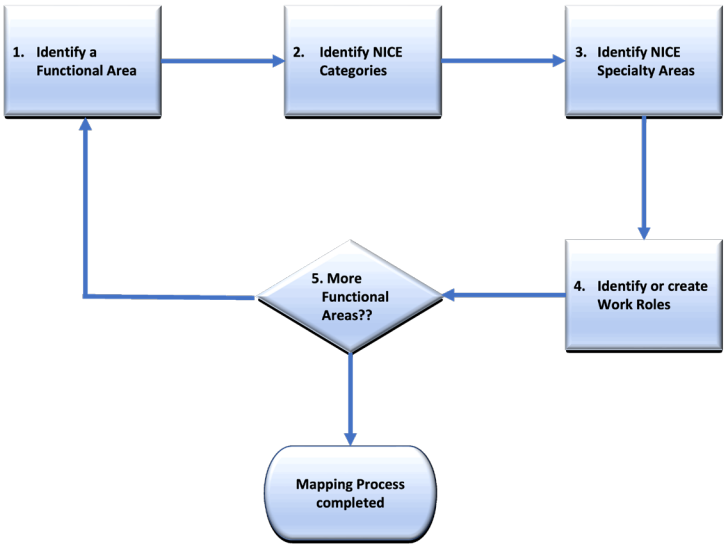


Figure 3: A Job Role mapping process

Step 1 – identify Functional Areas

This step aims to identify the CCI Competence Framework Functional Area and examine its description for mapping with the NICE Framework. However, due to the complexity of CCI and limited information regarding CCI Job Roles, in this step, we determine characteristics that typically identify each Dimension and each Functional Area for ease of reference during the mapping process. These characteristics are extracted from the description, functions, responsibility expectations and goal of each Dimension and its Functional Areas. Therefore, these characteristics are compared to the NICE Framework's Category and Specialty Area, which will be reflected in Step 2.

This paper limits the mapping process of identifying Job Roles to one Dimension and One Functional Area of the CCI Competence Framework. For example, we use Functional Area 5, Neutralisation under Dimension 4 – Active Offensive CCI (see Figure 1) to illustrate the practicality of the mapping steps. The characteristics are determined by the following:

a. The Offensive CCI is concerned with detecting and directly gathering intelligence about adversaries' covert, espionage, or cyber operations or deceiving and manipulating them. This can be done inter alia by creating honeypots containing files with misinformation (Lee, 2015)
b. The Active Offensive CCI's ultimate goal is to infiltrate, penetrate, exploit or counterattack the adversary's covert cyber operations. The adversary is likewise provided with disinformation and has its interpretation manipulated (Duvenage, et al., 2020a).
c. Neutralisation is about conducting CCI activities that render the adversary's cyber activities and capabilities ineffective, inactive, a failure, and collapse (Stech & Heckman, 2018). Neutralisation of adversary cyber activities can be achieved by "destruction, paralysis, loss of interest or loss of confidence that collection will be able to achieve its objective". Neutralisation proactively, in real-time or immediately after the initial attack, counteracts or destruct an adversary's cyber offensive activities.

As discussed in the first paragraph of this step, Figure 4 shows the identified characteristics for Functional Area 5, Neutralisation, and Dimension 4, Active Offensive CCI extracted from points (a), (b) and (c) above.

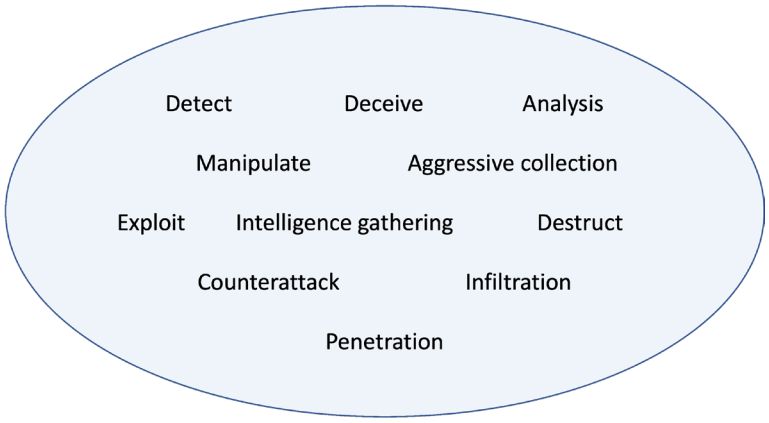


Figure 4: Example of characteristics for the Functional Area 5: Neutralisation [Dimension 4: Active Offensive CCI].

The following step identifies NICE Framework's Category for mapping with CCI Competence Framework's Functional Area.

Step 2 – identify NICE Categories

This step involves examining the descriptions of each of the seven NICE Framework Categories and then comparing them with the Functional Area characteristics identified in step 1 (see Figure 4). The outcome will be categories that have similarities with the characteristics and similar connotations of the descriptions of the Functional Area. For example, the result of this step is two NICE Framework Categories that correspond to the Functional Area 5, Neutralisation, of the CCI Competence Framework:

a. Analyse: performs highly specialised review and evaluation of incoming cybersecurity information to determine its usefulness for intelligence

b. Collect and Operate: provides specialised denial and deception operations and collection of cybersecurity information that may be used to develop intelligence.

The next step identifies the NICE Framework's Specialty Area to map with CCI Competence Framework's Functional Area.

Step 3 – identify NICE Specialty Areas

In this step,

a. We analyse and map every NICE Framework Specialty Area for each Category identified in Step 2 to the CCI Competence Framework's Functional Area identified in step 1.
b. Every Specialty Area that has similarities to Functional Area (based on the characteristics identified in Step 1) is selected. For example, the following seven Speciality Areas in Table 1 were identified for mapping Functional Area 5, Neutralisation, to every NICE Framework's Specialty Area.

Table 1: Identified NICE Framework Specialty Areas corresponding to the CCI Competence Framework Functional Area

CCI Competence Framework	NICE Framework Categories/Speciality Areas	
Dimensions/ Functional Areas	**Analyse**	**Collect and Operate**
Active-Offensive CCI: Neutralisation	All-Source Analysis Exploitation Analysis Targets Threat Analysis	Collection Operations Cyber Operational Planning Cyber Operations

Subsequently, the next step will identify relevant Job Roles from the NICE Framework Specialty Areas.s

Step 4 – identify or create Job Roles

This step involves two approaches:

a. Identify Job Roles from the identified Specialty Areas in Step 3 that are relevant to CCI Competence Framework. The mapping is based on examining whether the NICE Framework's Work Role is associated with the CCI Competence Framework's Functional Area, the CI/CCI functions (analysis, collection, investigation and operations) and CI/CCI goal/activities (identity, deter, prevent, destruct, manipulate, exploit and neutralising). It should be noted that the identified job role through this step is subjected to modification of job role name, code and description so that all CI and CCI taxonomy is reflected to appropriate the adaptation of the CCI Competence objective.
b. Create new Job Roles – this approach involves conducting further analysis of each CCI Competence Framework's Functional Area to determine missing responsibilities, functions or activities to reflect the purpose of the Functional Area and, in a broader sense, the mission of CCI. The creation of the new Job Roles is not yet conducted because they go beyond the scope of this paper.

For example, each Work Role from every identified NICE Framework's Specialty Area (Analyse, and Collect and Operate) is mapped to the CCI Competence Framework's Functional Area 5, Neutralisation, by examining whether the definition of the Work Role is associated with Neutralisation. The outcome of the mapping effort has identified nine relevant Job Roles, as shown in Table 2.

Table 2: Identified NICE Framework Job Roles for the CCI Competence Framework Functional Area Neutralisation (source: Newhouse, et al. (2017))

Category	Speciality Area	Work Role	Work Role description
Analyse	All-Source Analysis	Mission Assessment Specialist	Develops assessment plans and measures of performance/effectiveness. Conducts strategic and operational effectiveness assessments as required for cyber events. Determines whether systems performed as expected and provides input to the determination of operational effectiveness.
		All-Source Analyst	Analyses data/information from one or multiple sources to conduct preparation of the environment, respond to requests for information, and submit intelligence collection and production requirements in support of planning and operations.
	Exploitation Analysis	Exploitation Analyst	Collaborates to identify access and collection gaps that can be satisfied through cyber collection and/or preparation activities. Leverages all authorised resources and analytic techniques to penetrate targeted networks.

Category	Speciality Area	Work Role	Work Role description
	Targets	Target Network Analyst	Conducts advanced analysis of collection and open-source data to ensure target continuity; to profile targets and their activities; and develop techniques to gain more target information. Determines how targets communicate, move, operate and live based on knowledge of target technologies, digital networks, and the applications on them.
	Threat Analysis	Threat/Warning Analyst	Develops cyber indicators to maintain awareness of the status of the highly dynamic operating environment. Collects, processes, analyses, and disseminates cyber threat/warning assessments
Collect and Operate	Collection Operations	All Source-Collection Requirements Manager	Evaluates collection operations and develops effects-based collection requirements strategies using available sources and methods to improve collection. Develops, processes, validates, and coordinates submission of collection requirements. Evaluates performance of collection assets and collection operations
	Cyber Operational Planning	Cyber Intel Planner	Develops detailed intelligence plans to satisfy cyber operations requirements. Collaborates with cyber operations planners to identify, validate, and levy requirements for collection and analysis. Participates in targeting selection, validation, synchronisation, and execution of cyber actions. Synchronises intelligence activities to support organisation objectives in cyberspace
		Cyber Ops Planner	Develops detailed plans for the conduct or support of the applicable range of cyber operations through collaboration with other planners, operators and/or analysts. Participates in targeting selection, validation, synchronisation, and enables integration during the execution of cyber actions.
	Cyber Operations	Cyber Operator	Conducts collection, processing, and/or geolocation of systems to exploit, locate, and/or track targets of interest. Performs network navigation, tactical forensic analysis, and, when directed, executes on-net operations.

Based on the mapping results shown in Table 2, Figure 5 presents the nine identified Job Roles relevant for the CCI Competence Framework.

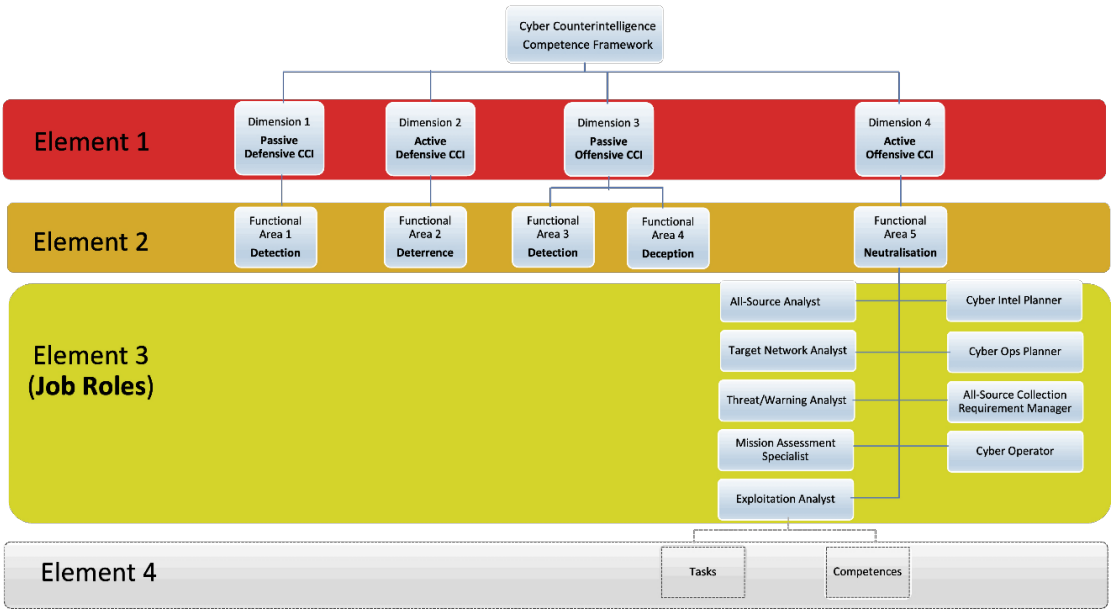


Figure 5: A CCI Competence Framework with Identified Job Roles (Element 3) for the Functional Area Neutralisation

Step 5 – More Functional Areas?

This step repeats the mapping process from Step 1, selecting the next Functional Areas until all Functional Areas are identified and have been associated with Job Roles.

This section presented a 5-step mapping approach for identifying Job Roles for the CCI Competence Framework.

For the purpose of this paper, the focus was solely on technical Job Roles and other critical determinants of CI and CCI. Owing to the comprehensive process, which goes beyond the scope of this paper, for job role identification and description, Functional Area Neutralisation was used as an example. The results can be

implicitly interpreted as a guideline for defining high-level identification, categorisation and definitions of Job Roles for the CCI Competence Framework.

4. Conclusion

This paper proposed a 5-step mapping process for identifying Job Roles, using a NICE Framework as a base framework. The mapping effort demonstrated the practicability of the relationship between some of the NICE Framework Categories / Specialty Areas and CCI Competence Framework Functional Areas. Although the paper was limited to identifying Job Roles for one Functional Area, nine Job Roles appropriately matched the Functional Area Neutralisation's purpose and functions. Each identified job role can be modified to fit the CCI discipline. The 5-step process allows for creating new Job Roles where Functional Areas' function and purpose are not adequately addressed. In completing the identification of Job Roles for the CCI Competence Framework, the following future work will need attention: (i) modification of identified Job Roles to align with the CCI-specific context, (ii) creation of new Job Roles to ensure that the CCI discipline a pool of requisite Job Roles to counter cyber-attacks, and (iii) for each identified and created job role, provide detailed information of the fundamental components of the job role (job title, job code, job purpose, tasks, level of proficiency and competences (knowledge, skills and abilities).

References

Check Point Research, 2023. *Check Point Research.* [Online] Available at: https://blog.checkpoint.com/2023/01/05/38-increase-in-2022-global-cyberattacks/ [Accessed 15 January 2023].

Duvenage, P. & von Solms, S., 2014. P*utting Counterintelligence in Cyber Counterintelligence.* University of Piraeus, Greece, 14th European Conference on Cyber Warfare and Security.

Duvenage, P. & von Solms, S., 2015. Cyber Counterintelligence: Back to the Future. *Journal of Information Warfare,* 13(4), pp. 42-56.

Duvenage, P., Jaquire, V. & von Solms, S., 2020a. Cyber Counterintelligence Matrix for Outsmarting Your Adversaries. *Journal of Information Warfare,* 19(1), pp. 1-11.

Furnell, S., 2021. The cybersecurity workforce and skills. *Computer & Security ,* Volume 100, p. 102080.

Guembea, B. et al., 2022. The Emerging Threat of Ai-driven Cyber Attacks: A Review. *Applied Artificial Intelligence,* 36,(1), pp. e2037254-2378.

Hajny, J. et al., 2021. Framework, Tools and Good Practices for Cybersecurity Curricula. *IEEE Access,* Volume 9, pp. 94723-94747.

Heckman, K. E. et al., 2011. Cyber Denial, Deception and Counter Deception A Framework for Supporting Active Cyber Defense. *Advances in Information Security,* Volume 64, p. 2011.

Jaber, A. & Fritsch, L., 2023. *Towards AI-powered Cybersecurity Attack Modeling with Simulation Tools: Review of Attack Simulators.* 249-257, International Conference on P2P, Parallel, Grid, Cloud and Internet Computing.

Lee, R. M., 2015. *Cyber Intelligence Part 4: Cyber Counterintelligence From Theory to Practices,* s.l.: Tripwire.

Microsoft, 2022. *Microsoft Digital Defense Report 2022 Illuminating the threat landscape and empowering a digital defense.,* s.l.: Microsoft.

Neal, A., Yeo, G., Koy, A. & Xiao, T., 2012. Predicting the form and direction of work role performance from the Big 5 model of personality traits. *Journal of Organizational Behavior,* Volume 33, pp. 175-192 .

Newhouse, W., Keith, S., Scribner, B. & Witte, G., 2017. *NIST Special Publication 800-181 National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework..* [Online] Available at: https://www.nist.gov/file/359276 [Accessed 2 July 2018].

Oxford Dictionary of English, 2010. *Oxford Dictionary of English,* s.l.: Oxford University Press.

Prunckun, H., 2019. *Counterintelligence Theory and Practice.* Lanham, Maryland: Rowman & Littlefield Publishing Group, Inc.

Saltz, J. S. & Grady, N. W., 2017. *The Ambiguity of Data Science Team Roles and the Need for a Data Science Workforce Framework.* Boston, MA, USA,, 2017 IEEE International Conference on Big Data (BIGDATA).

Sithole, T. & Du Toit, J., 2022. *A Cyber Counterintelligence Competence Framework.* Chester, Proceedings of the 21st European Conference on Cyber Warfare and Security ECCWS 2022.

SPARTA, 2020. *Strategic Programs for Advanced Research and Technology in Europe: D9.1 Cybersecurity Skills Framework.* [Online] Available at: https://www.sparta.eu/assets/deliverables/SPARTA-D9.1-Cybersecurity-skills-framework-PU-M12.pdf

Stech, F. J. & Heckman, K., 2018. Human Nature and Cyber Weaponry: Use of Denial and Deception in Cyber Counterintelligence. In: *Cyber Weaponry, Advanced Sciences and Technologies for Security Applications.* Cham: Springer, pp. 13-27.

Trend Micro, 2022. *A Trend Micro Survey Report The State of Industrial Cybersecurity Prevent supply disruptions and ensure resilient operations,* s.l.: Trend Micro .

World Economic Forum, 2022. *The Global Risks Report 2022, 17th Edition, Insight Report,* Geneva: World Economic Forum.

World Economic Forum, 2023. *The Global Risks Report 2023, 18th Edition, Insight Report,* Geneva: World Economic Forum.

Static Vulnerability Analysis Using Intermediate Representations: A Literature Review

Adam Spanier and William Mahoney
University of Nebraska at Omaha, Omaha, USA

aspanier@unomaha.edu
wmahoney@unoamah.edu

Abstract: Static Analysis (SA) in Cybersecurity is a practice aimed at detecting vulnerabilities within the source code of a program. Modern SA applications, though highly sophisticated, lack programming language agnostic generalization, instead requiring codebase specific implementations for each programming language. The manner in which SA is implemented today, though functional, requires significant man hours to develop and maintain, higher costs due to custom applications for each language, and creates inconsistencies in implementation from SA-tool to SA-tool. A source of programming language generalization occurs within compilers. During the compilation process, source code is converted into a grammatically consistent Intermediate Representation (IR) (e.g. LLVM-IR) before being converted to an output format. The grammatical consistencies provided by the IR theoretically allow the same program written in different languages to be analyzed using the same mechanism. By using the IRs of compiled programming languages as the codebase of SA practices, multiple programming languages can be encompassed by a single SA tool. To begin understanding the possibilities the combination of SA and IRs may reveal, this research presents the following outcomes: 1) a systematic literature search, 2) a literature review, and 3) the classification of existing work pertaining to SA practices using IRs. The results of the study indicate that generalized Static Analysis using the LLVM IR is already a common practice in all compilers, but that the extended use of the LLVM IR in Cybersecurity SA practices aimed at finding vulnerabilities in source code remains underdeveloped.

Keywords: Vulnerability Detection, LLVM, Compilers, Intermediate Representation Analysis

1. Introduction

Static Analysis (SA) is a software practice focused on detecting patterns and deficiencies contained within the source code of a program, without executing the code. (Thomson, 2021; Bodden, 2018; Schmeelk et al., 2015) Originally used for early development program optimization (Bodden, 2018), SA is now widely utilized as an effective means to detect common vulnerabilities and weaknesses in static code bases (Nachtigal et al., 2022). In practice, SA can take on many forms of varying complexities. Some tools are concise, aiming only to detect formatting issues in code (Nachtigal et al., 2022; Dlint, 2022; JavaScript, 2022). Others offer the ability find minor bugs by observing common patterns occurring in the code (Nachtigal et al., 2022; PMD, 2022; SpotBugs, 2022). More complex SA implementations sometimes claim full vulnerability discovery capabilities (Nachtigal et al., 2022; Find Security Bugs, 2022; SecurityCodeScan, 2022).

As a well understood knowledge domain, current SA implementations have reached a level of sophistication and functionality that early implementations couldn't attain (Bodden, 2018; Johnson et al., 2013). This is due to streamlined pattern detection algorithms and more efficient data structure architectures contained within modern SA tools (Bodden, 2018). Even with high levels of sophistication and complexity, the core benefit of SA lies in the fact that a code-base containing potential malware can be analyzed quickly without risking infection of the host system (Johnson et al., 2013). This low resource, high security functionality means SA is still used extensively in most cybersecurity analysis activities (Johnson et al., 2013).

Even with the sophistication afforded most modern SA tools, several issues arise to the practical implementation of SA in large scale applications. With the blinding pace of technological improvement, the increasing levels of complexity, and the overwhelming sizes of modern application code bases, SA, even with it's dearth of research and sophistication, will always struggle to keep up with rapidly expanding technologies (Bodden, 2018). Due to the direction connection between the vulnerabilities and weaknesses discovered in source code and the SA application designed to find them, each SA tools must be both coded by hand and coded specifically for a single language (Bodden, 2018; Johnson et al., 2013). These laborious development requirements require that large SA tools must be coded to handle new programming languages that may or may not be rapidly evolving as the tool is developed. This can make SA tools lag behind in terms of functionality. Exacerbating the issue is the general lack of usability for most SA tools (Johnson et al., 2013). While the tool developers try to keep apprised of code base changes and new vulnerabilities, the user interface generally suffers from tacked-on additions and difficult-to-understand processes (Johnson et al., 2013). Breaking through these barriers requires a new way of thinking in order to alleviate the code-specific nature of SA tools (Bodden, 2018).

1.1 Compilers

Compilers are small software packages that translate one programming language into another (Zhong et al., 2009; Aho et al., 1986). The input language is called the source language, while the output is called the target language (Zhong et al., 2009). In general, compilers convert a high level programming language like Fortran, C, C++, or Java into lower level executable machine code; however, compilers are not limited to high/low conversions and can also convert low-level code to high-level languages (Aho et al., 1986).

To carry out this translation process, compilers first analyze the input code and break it into smaller, constituent pieces. These constituent pieces are then passed through a two-phase analysis process wherein the lexical and syntactical characteristics of the input program are parsed to create an *intermediate representation* data structure, often in the form of a *Syntax Tree*. The resulting *Syntax Tree* allows the compiler to generate a program language agnostic, low-level representation of any given program, much like a program meant for an abstract machine. During the analysis portion of the program, should the input code present any number of deficiencies, the compiler generally responds using compiler errors. On the other side of the intermediate representation, the synthesis portion of the compiler is responsible for converting the internal intermediate representation into the desired target program.

1.2 Intermediate Representations

One of the most important elements of intermediate representations is that they must be fast and efficient to both produce and translate. Intermediate representations processes and structures are used in every compiler activity, thus, all intermediate representations must be fast and efficient. Any lack of efficiency will cause the functionality of the compiler to suffer.

It is important to note that the intermediate representation of each compiler exhibits consistent grammatical characteristics regardless of the input or output format. This consistency is necessitates so that a program written in C or Fortran to carry out the same computational steps will be compiled into identical source code outputs. This fundamental consistency within compiler intermediate representations can act as a bridge between languages, describing the functionality of the language in a language-independent grammar.

1.3 Intermediate Representations and Static Analysis

This language-independent, inter-compiler grammatical consistency, when coupled with SA practices can theoretically facilitate the design and creation of a SA tool that analyzes the intermediate representation of the compiler rather than specific high level programming languages. By combining these two elements, the burden of recoding SA tools for new or changing languages can be mitigated in favor of SA tools that specifically analyze the consistent intermediate representation. This research aims to explore foundational literature to gain a better understanding of the state-of-the-topic regarding SA using intermediate representations. This exploration will be achieved by meeting the following objectives: 1) describe the problem, 2) design a systematic literature review, 3) carry out a literature survey, 4) synthesize the findings, and 5) discuss the outcomes.

2. Survey Design

The literature survey conducted in this research follows processes set by (de Sousa Borges et al., 2014). The literature survey follows the following four steps:

1. **Research Questions**: The research questions in this study fall into the following classifications: 1) pertinent literature, and 2) classifications.
2. **Literature Search**: Based on the research questions developed, a replicable process is implemented to gather generally relevant literature.
3. **Literature Selection**: The literature collected is further refined by using a replicable process to select relevant material.
4. **Literature Synthesis**: The literature collected is synthesized and classified to determine patterns or existing themes.

2.1 Goals and Research Questions

The goal of this study aims to both discover and synthesize pertinent literature related to the use of intermediate representations in Static Analysis activities. To achieve this goal, the following research questions are asked:

1. What literature exists relating the use of intermediate representations and compilers to Static Analysis?
2. What common themes and characteristics occur in existing literature?
3. Can these commonalities be used to classify pertinent literature?
4. What themes, information, and knowledge does existing literature present about the use of intermediate representations in Static Analysis practices?

3. Literature Survey

While the application of SA using IRs presents new and novel opportunities, existing literature provides useful insights as to the current state of the practice. The work reviewed in this survey generally falls into three categories: 1) motivating literature, 2) the use of IR in general program analysis, and 3) the use of IRs in bug detection and error discovery.

3.1 Motivating Literature

By its very nature, code is complicated. Adding to this inherent complexity, the progression toward more complex and ubiquitous technological solutions makes understanding modern applications all the more difficult (Thomson, 2021). In such an environment, programmers need as many tools as they possibly can to help manage and understand complex code. Static analysis (SA) is one such tool. SA is defined as "the extraction of facts from another program's source code, without executing the program in question and usually as a distinct stage in the day-to-day software development process" (Thomson, 2021).

3.1.1 Static Analysis

Thomson (2021) indicates that SA can manifest in the following ways: 1) discovery of security vulnerabilities, 2) code beautification and standardization, 3) dead code detection, 4) taint analysis, 5) race condition detection, 6) bounds checking.

One of the most common types of SA lies in it's prolific use in compilers. Compilers break source code into pieces, pass these pieces through lexical and syntactic static analysis processes and convert the resulting syntax trees into low level *intermediate representations* (Aho et al., 1986). The lexical and syntactical processes help determine the existence of compiler errors, code formatting issues, syntactical errors, uninitialized variables, or suspicious variable assignments in the code (Thomson, 2021). In this way, the core function of a compiler serves as a means to extract facts about the source code, thus tying compilers intrinsically to the SA process. In general applications, compilers derive facts about the source code and manifest the existence of those facts as machine code.

Modern SA implementations generally manifest through code specific analysis applications generally based in Version Control, Continuous Integration-based, or IDE based environments (Thomson, 2021). Though SA is currently considered a separate step in software development, compilers and SA often end up working synchronously, sometimes to the point of integration into new versions of compiler software. This relation, however, can also cause an arms race between SA and compilers. As code-bases get updated, modified, and become evermore complex, SA must attempt to remain effective within the context of the dynamic code-base. In current practice, this manifests through laborious efforts on the part of the SA tool developers. Such an inefficient process necessitates newer and more efficient analysis as code-bases get bigger and more complex (Thomson, 2021).

3.2 IR in General Program Analysis

Most general SA activities in compilers, should they occur, happen at the intermediate representation level (IR). Due to the consistency in program logic, structure, and control, IRs can provide a language agnostic access point for programs of varying source code languages.

The use of IR in general program SA remains somewhat novel in application. A number of limited feasibility studies and experiments relating directly to general program analyses using IR are described below (Kataev et al., 2018; Belyaev et al., 2013; Ghime et al., 2022; Khaldi et al., 2016).

3.2.1 SAPFOR: Automatic Parallelization using IR

Kataev et al. (2018) provide insight into the state of the numerous and complex computational environments that exist today and the challenges that are posed by the automatic parallelization processes required to fit into these environments. With so much processing power now available, effective automatic parallelization of compiled programs via the use of compilers remains somewhat elusive.

As stated by Kataev, *System FOR Automated Parallelization* (SAPFOR) is a tool that can simplify the development of automatically parallelized programs. SAPFOR functions by carrying out source-to-source transformations such that a single, linear progressing program can be divided into portions that can be allotted to parallelized computing systems. This process generates a need to better represent and analyze programs originating from multiple high-level programming languages; notably, C and FORTRAN.

To meet this need, the research provided by presents a novel auto-parallelization application using SAPFOR, LLVM, and the LLVM-IR. The program outlined by first examines compiled programs via the LLVM-IR, then determines memory locations, and ensures that no two pointers generate homogeneous memory locations. This process is carried out by analyzing IR-level memory locations and mapping them to higher level source code and creating what the researchers call a source-level alias tree.

The outcomes of the research carried out by indicate that the novel source-level alias tree created in this work provides usefulness in automation of parallelization in C and FORTRAN-based programs using the LLVM architecture. The primary contribution of this work is the novel application of LLVM-IR analysis for more efficient, automatic parallelization operations.

3.2.2 Type and Effect Systems in SA using IR

Belyaev et al. (2013) outline a novel system that implements type and effect systems to analyze programs using their derivative LLVM-IRs. As a feasibility study, this paper serves as means to determine whether such a tool is possible via the exploration of general concepts and a novel design.

Belyaev et al. chose the LLVM-IR approach due to the need for a more consistent, language agnostic approach to static analysis. C and C++ programs can prove complex when analyzed in their source format. An IR accurately represents the program regardless of source language, while providing less complex representations of the structures involved. The researchers in this case used the LLVM default Static Single Assignment (SSA) representation model for their IR. The LLVM SSA IR is called bitcode, and forms the input of the novel system proposed by the researchers.

The outcomes of the research carried out by Belyaev et al. saw the creation of both a bitcode parser and a basic analysis library and the development of a proof of concept type and effect system that can detect undefined variables.

3.2.3 IR Mapping in Incremental Analysis

Ghime et al. (2022) posit that the use of intermediate representations (IR) in SA can provide useful insights as to the functionality of a given program. They indicate that, while the use of IR objects in one-pass SA is very useful, each resulting IR of a given program as it changes over time generates different IDs for each program entity. For example, a variable contained within the program could receive a different unique identifier each time the IR is modelled. This creates significant boundaries to tracking program entity SA information through time.

Incremental analysis is the study of programs as they evolve over time. Incremental analysis is generally carried out on the source code of programs as it changes during the System Development Lifecycle. While this more traditional approach functions well in a historic sense, it can be laborious, inefficient, and language specific. The use of IRs in incremental analysis can allow for a more streamlined process as a program is analyzed over time.

The research carried out by Ghime et al. presents an accurate approach by which IR objects can be mapped to each other through time, such that SA information can be carried forward throughout the development of a software package.

3.2.4 Automatic HBM Allocation using IR

Khaldi et al. (2016) present a novel means by which High Bandwidth memory (HBM) can be automatically allocated during the execution of a program. Much like the research carried out by Kataev et al. (2018), Khaldi

et al. (2016) seek to implement the use of IR SAs so that the optimizations that occur within a compiler can be translated into more efficient computation. Where Kataev et al. (2018) seek to balance processing loads between multiple environments, Khaldi et al. (2016) seek to optimize the usage of memory allocations.

According to, HBM systems utilize a new three-dimensional memory stacking process by which memory chips are stacked vertically using processor dies. This stacking process allows the memory chips to work together and act like one single device. By changing the architecture, HBM allows for much higher bandwidth performance. While HBM achieves high bandwidth, traditional DDR memory still operates at a much lower latency, therefore making IO much faster to and from the DDR modules. These balanced benefits necessitate the dynamic combination of both HBM and DDR memory-types during execution to achieve the best computational output.

To meet the highly dynamic requirements needed to automatically allocate program entities to either HBM or DDR, chose to utilize a compiler-based optimization function that analyzes the IR of a given program and chooses which memory allocations using the *malloc* function should be given to HBM and which would be better served in DDR.

The outcomes of the work carried out by saw the implementation of the novel IR-based BCDA process in the LLVM compiler, and the subsequent dynamic allocation of HBW/DDR memory during execution. The resulting efficiency of the system increased the speed of the executing program by up to 2.29 times that of the normally executed binary (Khaldi et al., 2016).

3.3 IR in Static Vulnerability Analysis

Though the use of IRs in program analysis relates directly to the questions asked in this research, the scope yet remains beyond the desired aim of this work. To hone in on IR and Static Vulnerability Analysis pairings, studies relating directly to bug detection using IRs are required.

Much like general program analysis using IRs, vulnerability analysis experiments using IRs are limited to a small number of application and feasibility studies (Liang et al., 2016; Cassez et al., 2017; Fornaia et al., 2019; Schilling et al., 2022). While each study discovered relates directly to the problem in question, all outcomes thus far are limited to the detection of, at most, three types of software bugs.

3.3.1 MLSA: Static Bug Detection using IR

Liang et al. (2016) state that static analysis falls into five (5) categories: 1) control flow analysis, 2) data flow analysis, 3) model checking, 4) taint analysis, and 5) symbolic execution. Control flow analysis deals with operators contained in instructions. Data flow analysis concerns the operands that contain data in a program. Model checking constructs an automaton based on bug behavior patterns and checks for the existence of the automaton within the program. Symbolic execution sees an analysis of execution paths based on symbolic variable values.

Liang et al. present a static analysis tool called MLSA that uses the LLVM-IR to support limited bug discovery in multiple C-based programming languages. Currently MLSA can detect three bugs: 1) a division by zero error, 2) pointer overflows, and 3) dead code.

The system proposed by Liang et al. (2016) operates by using both Clang and Dragon Egg compilers to generate an LLVM-IR file from a C-based source code file. The LLVM-IR files generated in this research contain three (3) main sections: modules, functions, and basic blocks. The proposed program then processes the instructions in the module according to the Control Flow Graph (CFG) generated by the IR, and the relationships between the basic blocks contained in the IR. A Z3 SMT solver is then used to determine if bugs exist within the IR model. When a bug is found, MLSA marks the location of the bug so it can be found quickly. These locations are then written into a log file.

3.3.2 Skink: Static Program Analysis using IRs

The research carried out by Cassez et al. (2017) aims to implement LLVM-IRs to determine the existence of error traces in the logical architecture of a given C-based program. To do this Cassez et al. generate an LLVM-IR using the Clang compiler. The resulting IR is then mapped onto a CFG. This graph is a Finite Labelled Automaton (FLA) that implements two specific labels: basic blocks and choices. The FLA accepts a regular language comprised of a set of traces that lead to a given error block. These traces, called abstract error traces, are not guaranteed to be feasible within the actual program. To check the correctness of a given program, Cassez et al. simply need to

determine if an abstract error trace is feasible in the IR CFG. Cassez et al. check feasibility by encoding a given abstract error trace into a logical statement and checking if the resultant output is satisfiable using an SMT. If satisfiable, an error trace has been discovered and the program can be called incorrect.

The outcomes of the research carried out by Cassez et al. is comprised of a software package called Skink. The package includes the full LLVM-IR suite, a front end, middle-end, and a back-end architecture. Skink has been shown to work on any programming language that can be compiled into LLVM-IR. The major benefits, as stated by Cassez et al., lie in the program's ability to discover loop invariants and the ability to establish program correctness.

3.3.3 JSCAN: IR-Based Static Analysis Framework

Fornaia et al. (2019) propose a cross-language framework that utilizes LLVM-IR for the analysis of Object-Oriented code. By using the language-agnostic nature of the LLVM-IRs generated by LLVM compatible languages, the tool developed, called JSCAN, provides a readily readable representation of IR contents. SA tools like BCEL can then be built on top of the generalized IR framework.

As stated by Fornaia et al., the main contribution JSCAN provides lies in it's ability to augment the LLVM-IR with a more nuanced and structured information support system. The two main characteristics of the JSCAN framework are: 1) it is easy to understand and work with, and 2) it is uniform to abstraction of a tool called BCEL.

The aim of the research carried out by Fornaia et al. lies in a desire to see other developers take the frameworks and build more comprehensive and diverse tools on top of the code developed by this research.

3.3.4 VANDALIR: IR-Based Vulnerability Analyses

According to Schilling et al. (2022), custom built static vulnerability analysis tools are expensive, time-consuming, and inefficient. Many man-hours are required to develop, code, and maintain functionality for very limited, very specific applications. Though SA tools can find hidden vulnerabilities and complex patterns in program structures, the implementation, maintenance, and upkeep can be both logistically inefficient and quite expensive. These limitations drive the need for newer and more dynamic methods by which static vulnerability analysis can be carried out on wider code bases in more simple contexts.

Schilling et al. present a novel approach by which Datalog programming is modelled as an LLVM-IR and subsequent static vulnerability activities are carried out on the new, intermediate representation. As an added bonus, the researchers also generate a new static analysis detection rule that helps detect stack based memory corruption in 90% of test cases in LLVM-IR programs.

Table 1: Literature Review Findings

Authors	Origin	Purpose	Type of Source	Themes	Classification
Thomson (2021)	USA	To explore difficulties and potential improvements that can be made to static analysis in cybersecurity	Discussion	Generalized Static Analysis, Static Analysis using Compilers	Motivating
Kataev (2018)	Armenia	An exploration of IR analysis for automatic parallelization during compilation	Research	IR Analysis, Parallelization	General Program Analysis
Belyaev et al. (2013)	Russia	Research on IR parsing and the creation of a basic IR analysis library for type and effect system variable analysis	Research	IR Parsing, IR Analysis, Type and Effect Systems	General Program Analysis
Ghime et al. (2022)	India	To explore the application of consistent labels to IRs such that incremental analysis can be undertaken.	Research	IR Parsing, IR Analysis, IR Labeling, IR Objects, Incremental Analysis	General Program Analysis
Khaldi et al. (2016)	USA	Explore novel means to distribute memory allocations between High Bandwidth Memory and DDR Memory using IR analysis to modify system calls during compilation	Research	IR Parsing, IR Analysis, Compilation Modification, High Bandwidth Memory	General Program Analysis
Liang et al. (2016)	China	To explore the discovery of bugs in source code via IR parsing and	Research	IR Parsing, IR Analysis, Vulnerability Detection	Static Vulnerability

Authors	Origin	Purpose	Type of Source	Themes	Classification
		analysis.			Analysis
Cassez et al. (2017)	USA	To investigate the use of IRs in determining source code error traces, loop invariants, and program correctness	Research	IR Parsing, IR Analysis, Error Traces, IR Objects, Program Correctness	Static Vulnerability Analysis
Fornaia et al. (2019)	Italy	An exploration of a cross-language framework for the creation of IR objects that can be used with cybersecurity static analysis tools.	Research	IR Parsing, IR Analysis, IR Objects, Analysis Framework, SA Tool Integration	Static Vulnerability Analysis
Schilling et al. (2022)	Italy	To explore a novel IR analysis method based on Datalog programming for vulnerability and error detection	Research	IR Modelling, IR Parsing, IR Analysis, Vulnerability Detection	Static Vulnerability Analysis

4. Themes and Classifications

As an answer to RQ1, the literature outlined in the preceding sections provides a general understanding of how IRs are being used in Static Analysis activities. To answer RQ2, the common themes and characteristics discovered during this review fall into three areas: 1) literature that provides impetus and motivation for IR-based static analysis, 2) literature regarding general LLVM IR-based static analysis, and 3) literature carried out in exploration of static vulnerability analysis using LLVM IRs.

As an answer to RQ3, the relevant works selected during the course of this study fall into three broad categories as defined by the themes stated above: 1) motivation, 2) general program analysis using IRs, and 3) vulnerability detection using IRs. These categories also represent the most common themes discovered in the analysis of the literature surveyed.

To answer RQ4, the following information provides useful insights as to the use of IRs in Static Analysis. Motivating works serves not only as the motive to carry out the work in question, but also as a means to connect the proverbial dots between SA and the use of IRs. The foundational premises outlined by Thomson et al. (Thomson, 2021) drive the need to experiment with more efficient and diverse SA tools, while simultaneously outlining the intrinsic and inherent connection between compilers and static analysis. In this work, one piece of literature fell within the motivating category.

Many of the works selected for inclusion relate directly to the use of IRs in general program analysis. While the intent of this work is to determine the state of research based on static vulnerability analysis using IRs, the use of IRs in novel ways in program analysis practices can provide useful insights for novel static vulnerability techniques. Works included in this classification indicate the use of an IR via static program analysis. This work analyzes four general program analysis applications that incorporate IRs in the analysis structure.

The remaining works included in this research fall into the final and most pertinent category, vulnerability detection using IR. The works included in this classification demonstrate, in some capacity, the use of an IR in static vulnerability analysis. This classification contains the remaining four works analyzed in this study.

5. Conclusion

The use of intermediate representations in static analysis as a means to detect and discover vulnerabilities in software remains a relatively under-developed research area. While several applications analyzed in this study delved into the vulnerability discovery via IRs, the overall state of research relating to vulnerability detection in IRs is generally limited. At the same time, the use of IRs in general static analysis through internal compiler functionality has been discovered to be a standard practice. All compilers parse the logical structure of compiled programs to determine the existence of compiler errors. This connection is what drives the need to explore the relationships between static analysis practices, compilers, and IRs. The extraction of static analysis data from IRs outside the compiler structure, however, is still relatively unexplored.

Through the limited observations put forward by this research, more effort should be placed into a deeper understanding of how IRs can better be implemented in static vulnerability analysis. Should the state of practice of IR-based SA find a greater level of sophistication, the potential for full vulnerability detection integration within many compilers could potentially occur, adding significant security robustness to modern compiler

architectures. By streamlining and integrating such a process, most compiled programs could account better for basic vulnerabilities, therefore increasing general security overall.

Through the limited observations put forward by this research, more effort should be placed into a deeper understanding of how IRs can better be implemented in static vulnerability analysis. Should the state of practice of IR-based SA find a greater level of sophistication, the potential for full vulnerability detection integration within many compilers could potentially occur, adding significant security robustness to modern compiler architectures. By streamlining and integrating such a process, most compiled programs could increasingly search for basic cybersecurity vulnerabilities, therefore increasing general security overall. Our research team is currently exploring ways in which the LLVM-IR can be utilized in this way; we hope to present research papers with our results at a future ECCWS or ICCWS venue.

References

Aho, A., Sethi, R., and Ullman, J. (1986). Compilers: Principles, Techniques, and Tools. Pearson Education Inc.

Belyaev, M. and Tsesko, V. (2013). LLVM-Based Static Analysis Tool Using Type And Effect Systems. Automatic Control and Computer Sciences. Volume 46.

Bodden, E. (2018). Self-Adaptive Static Analysis. Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results. ICSE-NIER ’18. New York, NY, USA: Association for Computing Machinery. 45–48.

Cassez, F., Sloane, A.M., Roberts, M., Pigram, M., Suvanpong, P., and de Aledo, P.G. (2017). Skink: Static Analysis of Programs in LLVM Intermediate Representation. International Conference on Tools and Algorithms for the Construction and Analysis of Systems. Springer. 380–384.

de Sousa Borges, S., Durelli, V.H., Reis, H.M., and Isotani, S. (2014). A Systematic Mapping on Gamification Applied to Education. Proceedings of the 29th annual ACM symposium on applied computing. 216–222.

Dlint. (2022). [Online]. Available: https://github.com/dlint-py/dlint

Find Security Bugs. (2022). [Online]. Available: https://find-sec-bugs.github.io/

Fornaia, A., Scafiti, S., and Tramontana, E. (2019). JSCAN: Designing An Easy to Use LLVM-Based Static Analysis Framework. 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). 237–242.

Ghime, V., Khadsare, A., Jana, A., and Chimdyalwar, B. (2022). IR Mapping: Intermediate Representation (IR) Based Mapping to Facilitate Incremental Static Analysis. 15th Innovations in Software Engineering Conference. 1–5.

JavaScript standard style. (2022). [Online]. Available: https://standardjs.com/

Johnson, B., Song, Y., Murphy-Hill, E., and Bowdidge, R. (2013). Why Don’t Software Developers Use Static Analysis Tools to Find Bugs? Proceedings of the 2013 International Conference on Software Engineering. ICSE ’13. IEEE Press. 672–681.

Kataev, N. (2018). Llvm-Based Approach to Static Analysis Of C Programs In SAPFOR. 19–23.

Khaldi, D. and Chapman, B. (2016). Towards Automatic HBM Allocation Using LLVM: A Case Study With Knights Landing. 2016 Third Workshop on the LLVM Compiler Infrastructure in HPC (LLVM-HPC). 12–20.

Liang, H., Wang, L,. Wu, D., and Xu, J. (2016). Mlsa: A Static Bugs Analysis Tool Based on LLVM IR. 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD). 407–412.

Nachtigall, M., Schlichtig, M., and Bodden, E. (2022) A Large-Scale Study of Usability Criteria Addressed by Static Analysis Tools. Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis. ISSTA ‘22. New York, NY, USA: Association for Computing Machinery. 532–543.

PMD. (2022). [Online]. Available: https://pmd.github.io/

Schilling, J. and Muller, T. (2022). Vandalir: Vulnerability Analyses Based on Datalog And LLVMiIR. Detection of Intrusions and Malware, and Vulnerability Assessment: 19th International Conference, DIMVA 2022, Cagliari, Italy. 96-115.

Schmeelk, S., Yang, J., and Aho, A. (2015) Android Malware Static Analysis Techniques. Proceedings of the 10th Annual Cyber and Information Security Research Conference. CISR ‘15. New York, NY, USA: Association for Computing Machinery.

SecurityCodeScan. (2022). [Online]. Available: https://security-code-scan.github.io/

SpotBugs. (2022). [Online]. Available: https://spotbugs.github.io/

Thomson, P. (2021). Static Analysis: An Introduction. The Fundamental Challenge of Software Engineering is One of Complexity. Volume 19, Number 4, 29–41.

Zhong, S., Shen, Y., and Hao, F. (2009). Tuning Compiler Optimization Options Via Simulated Annealing. 2009 Second International Conference on Future Information Technology and Management Engineering. 305–308.

Legal and Ethical Issues of pre-incident Forensic Analysis

Iain Sutherland[1], Matthew Bovee[2], Konstantinos Xynos[3] and Huw O. L. Read[2]
[1]Noroff University College, Tordenskjoldsgate 9, 4612 Kristiansand S, Norway.
[2]Norwich University, Northfield, VT, USA
[3]Mycenx Consultancy Services, Germany

iain.sutherland@noroff.no
mbovee@norwich.edu
kxynos@mycenx.com
hread@norwich.edu

Abstract: Investigators searching for digital evidence may encounter a variety of different IoT (Internet of Things) devices. Data in such devices and their environments can be both valuable, but also highly volatile. To meet best practices and to process these devices in an expeditious and forensically-sound manner, an investigator should have a predefined plan. Developing such plans requires prior knowledge developed through the exploration and experimentation of the "target" devices. The expanding variety, number, and pervasiveness of IoT devices means there is an increasing need for pre-incident analysis to ensure forensic tools and techniques acquire, preserve and document evidence appropriately. Many of these IoT devices have proprietary file- and operating-systems and may employ mechanisms to protect intellectual property by limiting or preventing access by researchers. Disassembly of the device and circumventing these mechanisms may be restricted by contract, end-user licence agreement (EULA) or legislation regarding intellectual-property rights. Legislative exclusions exist for security research, in some jurisdictions, permitting legitimate analyses. The pre-incident analyses of hardware to establish a forensic process bear some similarity to vulnerability and security research, however there are distinct differences in their end goals. This paper discusses the legal and ethical issues that may be encountered when conducting pre-incident forensics analyses focussing on IoT hardware. It highlights areas of particular concern, identifies best practice and subjects requiring future work as presented in the literature before providing a series of recommendations for forensics investigators processing these types of devices.

Keywords: Forensic practitioners, legal issues, ethical issues, vulnerability research, cyber forensics

1. Introduction: Current challenges

1.1 Challenges of analysis

IoT (Internet of Things) devices provide a variety of monitoring and control functions of personal and environmental factors. The data within these devices can be valuable for the digital investigator, yet highly volatile. Best practice requirements (NIST, 2019; ISO, 27037) demand validation and verification of forensic tools and processes for evidence collection before deployment at a crime scene or use on a device containing potential evidence. Investigators need a clear understanding of the impact of their actions on a device. The continued proliferation of IoT hardware means researchers should be engaged in "pre-incident" research. This can be defined as research focused on exploration of devices prior to an incident to: identify where evidence may be present prior to the device appearing in an investigation; determine how to acquire said evidence; and, document best practice for processing said device. Once an IoT device has been analysed, the results need to be disseminated and verified by others in the community. This verification process ensures that analyses and any proposed methods are correct and helps to identify exceptions, variations, or improvements. Pre-incident research is essential to support practitioners in present and future digital forensics investigations, but may have both ethical and legal implications.

1.2 Prior work

There has been previous work examining the legal and ethical issues of forensics research. Liles et al. (2009) explored the legal issues facing forensics experts which highlighted legal concerns including the interpretation of legislation. Read et al. (2015) focused on the legal and ethical issues of jailbreaking and hardware modification to access devices, and Stoykova et al. (2022) highlighted the legal and technical issues relating to the reverse engineering of file systems. Exploiting vulnerabilities in some cases provides the investigator access to devices that would otherwise be inaccessible. Pessolano et al. (2019) outlined a process for recovering Nintendo 3DS evidence that required use of a flashcart to exploit a vulnerability to access to the device. These flashcarts are advertised for illegally copying games and sale of these devices is now illegal in some jurisdictions. Other similar

examples of vulnerabilities used in investigations include checkm8 and checkra1n, used to exploit iOS devices by Cellebrite (Cellebrite, 2020).

In this paper we reiterate that digital forensics best practice requires an established process. Digital investigators therefore are faced with the need to explore a wide range of devices (i.e., using reverse engineering or other similar invasive techniques) as part of pre-incident analyses to expand their understanding of the device. Yet this process is time consuming, technically demanding, and presents legal and ethical challenges that need to be carefully managed.

The rest of this paper is organised as follows; Section 2 identifies the need for pre-incident analysis, Section 3 considers legal and ethical concerns, Section 4 considers the status quo and identifies areas in need of future effort. Finally, Section 5 provides a summary and conclusions.

2. The need for pre-incident analysis

2.1 The requirements of standards and best practice

Forensics researchers and practitioners are guided by standards and codes of best practice, with a key focus on the consistency and reliability of investigative practice and procedures to ensure the admissibility of evidence. These standards are clear on the need for established methodologies and for preparation prior to an incident. ISO/IEC 27037:2012, the international standard, provides guidelines for identification, collection, acquisition and preservation of digital evidence. In section 5.2 it states, "*All processes to be used by the DEFR* [Digital Evidence First Responder] *and DES* [Digital Evidence Specialist] *should have been validated prior to use*" (ISO/IEC 27037:2012). In addition, section 5.3.5 also highlights the importance of justification in handling, obtaining and reproducing digital evidence. It is therefore essential to test and validate a process prior to its use to ensure understanding of the device, and also to share processes and procedures for validation. This is further clarified in ISO/IEC 27041:2016 which outlines the steps required to ensure investigative processes are fit for purpose. This standard defines the way in which requirements can be captured and developed for the design and validation of any new investigative methods. It clearly states that a new method should be accompanied by a validation process with suitable test data.

2.2 Support for IoT Device Examination

Tools to support IoT investigations are limited. Manufacturers may be unwilling to provide device or data access, citing privacy or intellectual property concerns (SWGDE, 2022). A high-profile example that received much public attention and interest is Apple's reluctance to provide access to the contents of an iPhone in the Apple V FBI encryption case (United States District Court, 2016). A further challenge is that device types may also undergo rapid updates, so manufacturers might be *unable* to produce a duplicate device if the case in question involves an older or obsolete model. In addition, the device may have undocumented behaviour as a result of a vulnerability or modification. If manufacturers provide a new device this may exhibit different behaviour from an older evidentiary device that has been patched, updated and contains user data. Activities needed to aid understanding of a device may, in some cases, include the need to access and dump proprietary code for analysis (Dawson and Akinbi 2021). The limited support at present means that there may be the need to reverse engineer firmware to understand the operation of the device or to extract encryption keys to decrypt information held in the device (Pessolano et al. 2019).

2.3 Supporting Digital Forensics Practitioners

Device capabilities also need to be understood to support appropriate triage-process decision-making. Frameworks for IoT readiness and knowledge of IoT systems are highlighted as key areas for forensic readiness (Zulkippli et al., 2021). There is a need to understand the devices and systems if we are to address the types of problems mentioned by Luciano et al., (2018); there exists a

> *"... lack of funding towards research, training and forensic tools in IoT and these should be one of the most important research opportunities in the future ..."*

Horsman (2019) outlined that the pressures on the forensic practitioner are such that they may need to rely on the publication of forensic research to develop processes and procedures. There is a developing body of work exploring the expanding range of consumer devices to determine if the information left on systems might be of value to a forensic investigation and how to access this in a forensically sound manner. Examples include games

systems (Barr-Smith et al., 2021), home routers (Awasthi et al., 2018), smart TV's (Sutherland et al., 2014; Boztas et al., 2015) and robot vacuum cleaners (Zhou et al., 2022). Grouping such devices together under the banner of IoT is challenging due to varying levels of interactivity and security, copy protection mechanisms, and disparate types of evidence. Some work may even be funded by government research (e.g. Barr-Smith et al., 2021). However, researchers who share guidance in the forensic community to support investigations could come into conflict with manufacturers seeking to protect user data and intellectual property.

3. Legal and Ethical concerns

3.1 Legal issues

The focus on legal aspects of forensics research has been on examination and preservation of the evidence chain and on requirements of expert witness work (e.g. Huang, 2021). There appears to be a lack of discussion on the challenges surrounding the pre-incident analysis of systems as part of forensic readiness. Gamero-Garrido et al., (2017) examined the legal risks of third-party vulnerability research, noting it is possible to seek permission to carry out an analysis, but response rates varied between 20% and 40%. The work also noted 22% of security researchers polled had received legal threats due to their vulnerability research. It would be of interest to the digital forensics community to identify if there are similar issues when considering research supporting law enforcement, as is the focus of pre-incident analysis.

Beyond the "do" and "don't" approaches taken in End User Licence Agreements (EULAs), there are significant legal and best-practice documents that cover permissibility of security/vulnerability research. Table 1 presents several from the context of digital forensics research. Such legal protections broadly encompass pre-incident work to facilitate forensics on behalf of the legal system. Certain jurisdictions expressly exempt investigation action in support of law enforcement when backed by a court order. This would include forensics, although possibly only pursuant to an active case. Some jurisdictions allow for work in the "public interest", while other definitions of "good faith security research" do not appear to encompass digital forensics research, being more focussed on vulnerability and cyber security research aimed at protecting specific systems.

In some jurisdictions, such as the UK, there is no statutory defence for legitimate security research. As outlined in Table 1 the Crown Prosecution Service guidance (Crown Prosecution Service, 2020) states that various factors must be considered to determine if the prosecution is in the public interest, for instance; financial gain, concealment of identity and in particular how any tools or 'articles' created by the work have been used. The guidance states*; "For example, whether the article was circulated to a closed and vetted list of IT security professionals or was posted openly."* (Crown Prosecution Service, 2020) which has significant implications on the way that tools developed by academics as part of pre-incident research might be published, the how results of a pre-incident analysis can be validated, and whether any tools and datasets developed can be shared.

Table 1. Key Details of Example Pertinent Laws, Regulations, Acts, and Guidance.

Digital Millennium Copyright Act (1986)	Allows lawfully authorised investigative, protective, information security, or intelligence activity by officers, agents, or employees of the United States, a State, or a political subdivision of a State, or a person contracted by them. Other individuals are limited to a definition of "security testing" that does not appear to include research to identify forensically-sound means of identifying and acquiring evidence.
Computer Misuse Act (1990)	UK legislation with penalties of imprisonment and fines for any attempt (whether successful or even possible) to secure or facilitate unauthorised access to data. Although Crown Prosecution Service (2020) CMA Prosecution Guidance states where CMA violations are performed in the "public interest", prosecutors are advised to consider financial, reputational, or commercial damage caused to victims, and whether the purpose of the offence was for financial gain.
Computer Fraud and Abuse Act Charging Policy Guidelines (2022)	Prosecution should be declined if a defendant's conduct and intent was that of "good-faith security research". Such research is, however, narrowly defined in such a way that it does not appear to include digital forensics.
Electronic Frontiers Foundation (2021)	Electronic Frontier Foundation (EFF) Coders' Rights Project Vulnerability Reporting FAQ that provides guidance that focuses primarily on vulnerability research, but many of the recommendations and advisements apply to digital forensics, such as: seeking legal advice in advance; publishing only non-functional/non-code details/results; being aware of the impacts of copyright and trade-secret laws
Extraction of Information from electronic devices: code of practice, UK Home Office	Guidance refers to the Data Protection Act (DPA) and the General Data Protection Regulation (GDPR), both of which focus on protecting the data of

(Home Office 2022)	individual persons.
Council of Europe, Convention on Cybercrime,(2001)	Council of Europe, Convention on Cybercrime, European Treaty Series – No. 185 recommends each signatory adopt legislation and other measures to criminalise intentional unauthorised access to any or all of a computer system, or making available a "device" to aid such access.
Computer Fraud and Abuse Act (CFAA) (1986).	Sets a test value of USD 5,000 damage due to unauthorised access to any protected computer in any one-year period, and provides for imprisonment of up to five years

Guidance in relation to prosecution under the US Computer Fraud and Abuse Act (Computer Fraud and Abuse Act Charging Policy Guidelines, 2022) included good faith security research for the first time in 2022. The phrase 'investigation' is cited as a valid reason for research, but this is then followed with a statement that the primary goal of such work should be to improve the safety and security of the device, the services it supports, or the related users (Computer Fraud and Abuse Act Charging Policy Guidelines (2022). This appears not to include pre-incident analysis in support of a forensic investigation as a valid reason. Even if that *was* intended, there is nothing to preclude a vendor that believed it had been harmed by such investigative actions from filing suit and leaving it to the courts to evaluate the guidance.

3.2 Ethical issues

Pre-incident analysis also raises some issues from an ethics perspective. While various codes of ethics exist (Table 2), there is no unified code of ethics for digital forensics. The need for a code of ethics for digital investigations was highlighted by Seigfried-Spellar et al., (2017), and work such as that conducted by Horsman (2022) has explored the ethical issues and concerns from the privacy perspective and the impact of an investigation on the individual. Ferguson et al. (2020) in their paper highlight the ethical issues of an investigation on both individual privacy and organisational confidentiality. The issue of disclosing organisational intellectual capital / intellectual property is addressed, with the observation that forensics investigations have the possibility to disclose intellectual capital and therefore cause damage to an organisation.

A comparison of ethics for different professional organisations is presented in Table 2. Codes of practice, of ethics, and mixed aspects of both exist, but do little to resolve tension between the need for pre-incident examination, preparation and various legal constraints. Some codes say to act legally or the equivalent, while simultaneously exhorting activities (e.g., training, experience, validation of methodology) that benefit from pre-investigative research and preparation. The need, however, is to have broader documented exemptions and protections for work completed ethically that ultimately benefits the courts, Law Enforcement Agency (LEA) agencies, the legal system, their related processes and the community as a whole.

Table 2: Example Code Details from Select Organisations.

ACM Ethics (2023)	Section 2.3 states "Know and respect existing rules pertaining to professional work. where "Rules" here include local, regional, national, and international laws and regulations.
Digital Forensics Certification Board (DFCB) (2003)	Under "Professional Care and Competence": understand the nature and scope of work presented; be competent and accept engagements within one's scope of competence; and, utilise validated/appropriate methods, techniques, standards, and controls.
IEEE Code of Ethics (2003)	Improve the understanding by individuals and society of the capabilities and societal implications of conventional and emerging technologies; avoid unlawful conduct in professional activities; maintain and improve technical competence; undertake technological tasks for others only if qualified by training or experience, or after full disclosure of pertinent limitations; and, avoid injuring the property of others
International Association of Computer Investigative Specialists (IACIS) (2019)	Advise/assist within the scope of one's "legal authority"; be honest and ethical *when dealing with each other*
International Society of Forensic Computer Examiners (2009)	Never reveal anything confidential without a competent court order or express permission by the client, engage in any unethical or illegal conduct, or knowingly undertake an assignment beyond one's skills
(ISC)² (2023)	Act honourably, honestly, justly, responsibly, and legally (including observing all contracts and agreements, express or implied). Simultaneously, provide diligent and competent service to principals, and advance and protect the profession.

3.3 Legal and Ethical issues: A comparison to Vulnerability Research

The current ethical guidance and statutory defences present in legislation are focussed on narrowly-defined legitimate cyber security research, and typically relate to vulnerability analysis. Vulnerability research is

concerned with determining if security measures can be circumvented due to errors in the device's design or functionality. However, forensics is concerned with the information available in the device, and the accuracy of methods used to extract data. To be forensically useful, extractions should be repeatable, each iteration providing exact copies of the data held in the device.

In pre-incident research a vulnerability may be used as a step in the process. Video-game hacking, on which some forensics methods rely (Pessolano et al., 2019) shares similarities to vulnerability researchers, searching for issues in software and hardware, but as a means to obtain additional access to a games system. Device manufacturers apply various methods and techniques, like cryptography and obfuscation, as a defence mechanism to keep the system and its contents hidden from those wanting to engage in hacking/reverse-engineering of copy-protection mechanisms. The research performed on these proprietary systems is not dissimilar to "Blackbox" analysis. This requires extensive knowledge, experience and specialist tools, but has been known to generate results containing misunderstandings, errors and even omissions during the reverse engineering process. A digital forensic analyst relying on research from such groups must be wary of the temporal nature that afflicts these communities. Websites hosting information may disappear without notice, resellers of hardware that can circumvent security may go out of business (or may be removed because of legal proceedings), toolsets may not be updated and malfunction on newer versions of devices. All these issues create uncertainty and an almost ephemeral nature to digital forensic capability.

Responsible vulnerability research has a defined best practice of disclosing the issue to the manufacturers to enable a corrective action, a patch to be developed, before the vulnerability is disclosed to the wider community. Ideally in a pre-incident analysis to ensure the validation, it is the process, methods and tools developed that should be disseminated for validation. Horsman (2019) summarised the key attributes of an digital artefact that should be communicated for research to be of value to the forensics community including: the full details of the artefact; any methodology used; and, appropriate test data. The manufacturer is not involved in this process. Currently there are different approaches to this dissemination with some entities providing data for wider analysis (VTO Labs, 2023) and some such at the Artefact Gnome Project (Artefact Gnome Project, 2023) being more restricted.

In some cases pre-incident research may be a breach of the terms and conditions, and involve processes that may conflict with ethical codes of practice. It may also skirt close to what is permitted by national laws. Care must be taken to avoid publishing material that may be seen as promoting breach of copyright or the disclosure or theft of intellectual property.

4. Discussion and areas for future work

4.1 The need for legal clarification

Local legislation needs to address the issue of limitations introduced when making reverse engineering acts illegal. It is unclear what happens when law enforcement needs or relies on work products potentially deemed illegal or impermissible by local legislation. Legislation could also address concerns over manufacturers providing access to devices to circumvent encryption / privacy mechanisms. This is still a current concern for law enforcement (Wray, 2021).

Differences in legal and ethical frameworks could hamper forensics research in certain areas or jurisdictions, leading to an overreliance on material produced in *other* jurisdictions and loss of local capability. A valid concern for technology companies and lawmakers in changing legislation or in allowing further exemptions (such as those applied to the DMCA in the USA) would be the risk of accidentally permitting research that exposes intellectual property under the guise of valid forensics research. This would require some degree of regulation.

4.2 Clarification of codes of ethics

While there are numerous examples of codes of ethics, they have the challenge of addressing the tension between the absolute, strict interpretation of the legal constraints, the due diligence requirements of many codes of ethics/best-practice and of digital evidence examiner professions, the evidential needs of the legal system, and LEA and examiner resource limitations. It is clear that pre-incident analysis will require a strong set of ethical statements. Standard operating procedures and clear Code of Ethics are essential to define best practice and avoid unethical behaviour in this type of research. The authors would recommend one that is concise. The IEEE Code of Ethics (2023) provides a good example of such an attempt. We present three headings:

- To uphold the highest standards of integrity, responsible behaviour, and ethical conduct in professional activities.
- To treat all persons fairly and with respect, to avoid harassment or discrimination, and to avoid injuring others.
- To strive to ensure this code is upheld by colleagues and co-workers.

These provide appropriate high-level principles. In addition, the best practice guidance in Table 3 should be considered.

Table 3: Best Practices, also drawing on vulnerability research recommendations from the Electronic Frontiers Foundation (Electronic Frontiers Foundation, 2021).

Legal Process	**Legal within the jurisdiction** While a basic requirement, this is highlighted in *The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research* (Bailey, 2012) as a specific concern as *compliance* with the laws in that jurisdiction is seen as critically important.
Logical Device Selection	**The selection of the target of the pre-incident investigation should be described, appropriate and based on potential evidential value.** Any examination of a device should have a rational methodology for device selection based on type, availability, cost, potential evidence etc. The selection rationale should be open and clearly described.
Owned Device	**All experimentation should take place on a device purchased for the purpose.** Second-hand devices should be avoided (or blindly sourced) to avoid targeted recovery of personal data. If there is a risk of personal data being recovered then appropriate institutional ethical review may be needed.
Focused Analysis	**Although examination of a device might expose Intellectual Property, the analysis and results should focus on evidence sources and user activity / interaction. This should avoid proprietary areas / information where possible.** This should be focussed on potential evidence. This is usually user activity / settings / usage of the device etc. Care should be taken to protect Intellectual Property (IP) and not provide tools or techniques that access/expose areas of the device without need.
Limited Dissemination	**While peer reviewed publication is prevalent, and there is a shift for open access to all datasets and tools, for pre-incident analysis this should only be supplied to LE or other researchers for validation / verification and peer review.** The scientific approach requires verification / validation by other researchers. Dissemination should be limited to ensure access to tools and datasets only for validation / verification of the method, further research, and law enforcement use. Wider dissemination should be limited, if it would cause harm to the industry. Therefore, researchers would need to apply some mechanism to decide on the benefit to society.

Proactive forensics research needs to be done with care to ensure legal and ethical considerations are taken into account when deciding to examine a device. The permitted activities can depend very much on the jurisdiction in which the device is being examined. This can be an added complexity if the research team is international and spanning several different jurisdictions. One way to address ethical considerations may be to expand the oversight of Research Ethics Committees (RECs) (Europe) and Institutional Review Boards (IRB) (in the USA) which usually are focused on potential harm to the individual and attentive to personal rights as espoused in the Belmont Principles which address respect for persons, beneficence and justice (fairness and equality). These committees or boards could assess similar ethical considerations with regard to research, research processes and their results.

5. Summary and conclusion

International standards and best practice guidelines define the requirement for tested, verified methods for the forensic analysis of IoT devices. There is clearly a current and increasing need for forensics research to explore IoT devices prior to an incident occurring, to determine if information of evidential value is present, and to develop best practice for accessing and extracting that information. This forensic intelligence analysis clearly aligns with practices seen in the numerous on-going papers by researchers exploring devices. Labs and forensic practitioners need access to the research in published work (i.e., academic or blogs), but should also be able to access datasets to verify and better understand the processes presented.

There is a demonstrable need for pre-incident analysis, but also a clear tension with current legislation in some jurisdictions. Security and vulnerability research is already very familiar with this tension, although exemptions and codes of practice for prosecutors recognise the value of security-related research. The question remains whether ongoing work with pre-incident analysis falls into this category by design, by accident, or at all as some legislation clearly emphasises that the work is only permitted for the identification and reporting of security

vulnerabilities. There is a reliance on the current interpretation of procedural practice continuing to recognise the value of on-going forensic work. Perhaps the root of the issue, at least in the UK and USA, is that computing legislation is over 30 years old and technology and the associated law enforcement challenges have moved on considerably since this legislation was enacted. It is recognised that prosecution procedural practises are regularly updated, but this still poses a challenge, highlighted in the UK by the current campaign for a statutory defence for responsible security research to be enshrined in law (CyberUp, 2022).

In terms of ethical codes there is also a disparity between the need for pre-incident research on the phrasing of some current ethical codes. In this paper, current legal, ethical, and best practice for pre-incident digital forensics research has been discussed, and found to be lacking when compared to the more-established areas of vulnerability research. Opportunities for further work in the harmonisation between these two related areas, particularly in legal and ethical guidance is needed to ensure digital forensic capability is maintained in the future.

6. References

ACM Ethics (2023), ACM Code of Ethics and Professional Conduct, [Online] https://ethics.acm.org/

Artifact Gnome Project (2023) https://agp.newhaven.edu/about/start/

Awasthi, A.; Read, H.O.; Xynos, K.; Sutherland, I. (2018) Welcome pwn: Almond smart home hub forensics. Digital Investigation. 2018, 26, 38–46.

Bailey M., Dittrich D., Kenneally E., Maughan D. (2012) The Menlo Report. IEEE Security and Privacy Volume 10 Issue 2 March 2012 pp 71–75 https://doi.org/10.1109/MSP.2012.52

Barr-Smith F., Farrant T., Leonard-Lagarde B., Rigby D., Rigby S., Sibley-Calder F. (2021) Dead Man's Switch: Forensic Autopsy of the Nintendo Switch. Forensic Science International: Digital Investigation, Volume 36, Supplement, April 2021, 301110. https://doi.org/10.1016/j.fsidi.2021.301110

Boztas A, Riethoven A.R.J Roeloffs M. (2015) Smart TV forensics: Digital traces on televisions, Digital Investigation, Volume 12, Supplement 1, 2015,Pages S72-S80,ISSN 1742-2876, https://doi.org/10.1016/j.diin.2015.01.012

Cellebrite (2020) A Practical Guide to checkm8 – Cellebrite UFED. [Online] https://cellebrite.com/en/a-practical-guide-to-checkm8/

Computer Fraud and Abuse Act (CFAA) (1986). [Online] https://www.congress.gov/bill/99th-congress/house-bill/4718

Computer Misuse Act (1990) [Online] https://www.legislation.gov.uk/ukpga/1990/18

Computer Fraud and Abuse Act Charging Policy Guidelines (2022) [Online] https://www.justice.gov/jm/jm-9-48000-computer-fraud

Crown Prosecution Service (2020) Computer Misuse Act: prosecution guidance. [Online] https://www.cps.gov.uk/legal-guidance/computer-misuse-act

Council of Europe, Convention on Cybercrime,(2001) European Treaty Series – No. 185. (2001) [Online] https://rm.coe.int/1680081561

CyberUp (2022) Legitimate cyber security activities in the 21st century: Assessing the current consensus of what should constitute legitimate cyber security activity under a reformed UK Computer Misuse Act 1990. [Online] https://www.cyberupcampaign.com/s/CyberUp-campaign-V3-2022.pdf

Dawson, L., Akinbi, A., (2021) Challenges and opportunities for wearable IoT forensics: TomTom Spark 3 as a case study, Forensic Science International: Reports, Volume 3, 2021, 100198, ISSN 2665-9107, https://doi.org/10.1016/j.fsir.2021.100198.

Digital MIllennium Copyright Act (1998) [Online] https://www.copyright.gov/dmca/

Digital Forensics Certification Board (2003) Code of Ethics and Standards of Professional Conduct. [Online] https://dfcb.org/code-of-ethics-and-standards-of-professional-conduct/

Electronic Frontiers Foundation (2021) Coders' Rights Project Vulnerability Reporting FAQ [Online] https://www.eff.org/issues/coders/vulnerability-reporting-faq

Ferguson, R.I., Renaud, K., Wilford, S. and Irons, A. (2020), "PRECEPT: a framework for ethical digital forensics investigations", Journal of Intellectual Capital, Vol. 21 No. 2, pp. 257-290. https://doi.org/10.1108/JIC-05-2019-0097

Gamero-Garrido A., Savage S., Levchenko K., Snoeren A. C. (2017) Quantifying the Pressure of Legal Risks on Third-party Vulnerability Research. CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security October 2017 Pages 1501–1513 https://doi.org/10.1145/3133956.3134047

Home Office (2022) Extraction of Information from electronic devices: code of practice [Online] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1110883/E02802691_Electronic_Devices_Code_of_Practice_WEB.pdf

Horsman, G. (2019) Raiders of the lost artefacts: Championing the need for digital forensics research, Forensic Science International: Reports, Volume 1, 2019, 100003, ISSN 2665-9107, https://doi.org/10.1016/j.fsir.2019.100003.

Horsman, G. (2022) Defining principles for preserving privacy in digital forensic examinations, Forensic Science International: Digital Investigation, Volume 40, 2022, 301350, ISSN 2666-2817, https://doi.org/10.1016/j.fsidi.2022.301350.

Huang J., (2021) When Digital Forensic Research Meets Laws, 2012, 32nd International Conference on Distributed Computing Systems Workshops. https://doi.org/10.1109/ICDCSW.2012.45

IEEE Code of Ethics (2023) [Online] https://www.ieee.org/content/dam/ieee-org/ieee/web/org/about/corporate/ieee-code-of-ethics.pdf

International Association of Computer Investigative Specialists (2019) IACIS Code of Ethics and Professional Conduct. [Online] https://www.iacis.com/wp-content/uploads/2019/11/IACIS-Code-of-Ethics-and-Professional-Conduct-Ver-1.4.pdf

International Society of Forensic Computer Examiners (2009) Code of Ethics and Professional Responsibility, [Online] https://www.certified-computer-examiner.com/ethics2.htm

(ISC)² (2023) Code Of Ethics [Online] https://www.isc2.org/ethics/

ISO/IEC 27037:2012 Information technology — Security techniques — Guidelines for identification, collection, acquisition and preservation of digital evidence. [Online] https://www.iso.org/standard/44381.html

ISO/IEC 27041:2015 - Information technology — Security techniques — Guidance on assuring suitability and adequacy of incident investigative method. [Online] https://www.iso.org/standard/44405.html

Liles, S., Rogers, M., Hoebich, M. (2009). A Survey of the Legal Issues Facing Digital Forensic Experts. In: Peterson, G., Shenoi, S. (eds) Advances in Digital Forensics V. Digital Forensics 2009. IFIP Advances in Information and Communication Technology, vol 306. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04155-6_20

Luciano L., Baggili I., Topor M., Casey P., Britinger F., (2018) Digital Forensics in the Next Five Years. ARES 2018: Proceedings of the 13th International Conference on Availability, Reliability and Security August 2018 Article No.: 46 Pages 1–14 https://doi.org/10.1145/3230833.3232813

NIST (2019) Computer Forensics Tool Testing Program (CFTT) [Online] https://www.nist.gov/itl/ssd/software-quality-group/computer-forensics-tool-testing-program-cftt

Pessolano G., Read H.O.L., Sutherland I., Xynos K., (2019) Forensic Analysis of the Nintendo 3DS NAND, DFRWS-USA July 14-17 Portland, Oregon, USA in Digital Investigation: The International Journal of Digital Forensics & Incident ResponseVolume 29 Issue S Jul 2019 pp S61–S706 https://doi.org/10.1016/j.diin.2019.04.015

Read H., Sutherland I., Xynos K. and Roarson F., (2015), Locking out the Investigator: The need to circumvent security in embedded systems., Information Security Journal: A Global Perspective, vol24. Issues 1-3, 2015. Pages 1-9, http://dx.doi.org/10.1080/19393555.2014.998847

Seigfried-Spellar, Kathryn C.; Rogers, Marcus; and Crimmins, Danielle M.,(2017) "Development of A Professional Code of Ethics in Digital Forensics" (2017). Annual ADFSL Conference on Digital Forensics, Security and Law. 12.

Stoykova, R., Nordvik,R., Ahmed, M., Franke, K., Axelsson, S., Toolan,F., (2022) Legal and technical questions of file system reverse engineering,Computer Law & Security Review, Volume 46, 2022, 105725, ISSN 0267-3649, https://doi.org/10.1016/j.clsr.2022.105725

Sutherland I., Xynos, K., Read, H., Jones, A., Drange T., (2014) A forensic overview of the LG Smart TV, presented at the 12th Australian Digital Forensics Conference 2014 SRI Security Congress, "Security on the Move" 1-3 December, 2014, Perth, Western Australia.

Scientific Working Group on Digital Evidence SWGDE (2022) Best Practices for On-Scene Identification, Seizure, and Preservation of Internet of Things (IoT) Devices (2022)

United States District Court (2016) In the Matter of the Search of an Apple iPhone Seized During the Execution of a Search Warrant on a Black Lexus IS300, California License Plate 35KGD203, Case No. 15-0451M. [Online] https://www.documentcloud.org/documents/2722199-5-15-MJ-00451-SP-USA-v-Black-Lexus-IS300.html

VTO Labs(2023) IoT Forensics [Online] https://www.vtolabs.com/iot-forensics

Wray C.A. (2021) Statement of Christopher A. Wray Director Federal Bureau of Investigation before the Committee on the Judiciary United States Senate, at a hearing entitled “Oversight Of The Federal Bureau Of Investigation: The January 6th Insurrection, Domestic Terrorism, And Other Threats” Presented March 2, 2021. [Online] https://www.justice.gov/file/1525301/download

Zhou, H., Deng, L., Xu, W., Yu, W., Dehlinger J., and Chakraborty S., "Towards Internet of Things (IoT) Forensics Analysis on Intelligent Robot Vacuum Systems," *2022 IEEE/ACIS 20th International Conference on Software Engineering Research, Management and Applications (SERA)*, Las Vegas, NV, USA, 2022, pp. 91-98, https://doi.org/10.1109/SERA54885.2022.9806735

Zulkipli N.H.N., Willis G.B. (2021) An Exploratory Study on Readiness Framework in IoT Forensics Procedia Computer Science Volume 179, 2021, Pages 966-973 https://doi.org/10.1016/j.procs.2021.01.086

Developing Robust Cyber Warfare Capabilities for the African Battlespace

Mphahlela Thaba[1] and Jabu Mtsweni[1,2]
[1] Council of Scientific and Industrial Research, Pretoria, South Africa
[2] Security Institute for Governance and Leadership in Africa (SIGLA), Faculty of Military Science, Stellenbosch University

jthaba@csir.co.za
mtswenij@gmail.com

Abstract: The evolution of technology in the African battlespace continues to pose a significant challenge to the African militaries. This evolution increases the need for the African militaries to be able to operate in the cyberspace strategically and effectively. Developing cyber warfare capabilities remains a challenge to many African militaries who are struggling to remain afloat due to ever decreasing resources, including budgets. This in turn reduces the effect of these militaries in the evolving battlespace. This paper seeks to present a comprehensive framework for developing cyber warfare capabilities for African militaries to be able to operate efficiently in the cyber battlespace. The proposed POSTEDFIT aligned framework, requires a comprehensive system thinking approach towards developing capabilities in a phased manner. This includes the ability to define the capabilities in terms of the requirements presented by the cyberspace, and the components forming these capabilities. The generic framework is based on the basic understanding of a capability, as the ability to do something, in this case, the ability to secure and operate in the cyberspace for African militaries, ability to conduct offensive cyber operations and ability to keep abreast with the evolving cyber battlespace.

Keywords: Cyber Security, Cyber Warfare, Cyber Operations, Cyber Defence, Cyber Attacks, Africa Battlespace

1. Introduction

The interest in cyber warfare is demonstrated in various ways, and we are seeing other countries collaborating with their allies in developing cyber warfare capabilities. According to Ndebele (2023), the US Army has pledged to support selected African countries to improve their defences against extremists and terrorism in the cyber space, maritime, and land. Some of the countries targeted by this initiative include Botswana, Rwanda, Kenya, and others.

Nevertheless, there are still gaps in cyber warfare capabilities for most African countries. In fact, African countries are lagging far behind that only one African country appears in the top 30 of countries with the higher National Cyber Power Index (i.e., Egypt) (Voo, Hemani, & Cassidy, 2022). In addition, the African countries do not have established doctrines on cyber operations, lack of skills and capacity are a serious impediment, and reliance on foreign technologies is still a pain point for most African countries.

This paper therefore seeks to present a comprehensive framework for developing cyber warfare capabilities to enable African militaries to operate in the cyberspace, either to secure and/or to exploit for advancing military interests.

The rest of this paper is structured as follows: Section 2 provides context and overview on the definitions of terms used in this paper. In Section **Error! Reference source not found.**, the importance of cyber warfare in modern conflicts is narrated including the status quo in the African context. Section 4 highlights the research methodology adopted for this study in conducting the literature review as well as in conceptualising the proposed generic cyber warfare capability development framework. In Section 5, the different approaches that are used in cyber operations are demonstrated and discussed. Section 6, the proposed comprehensive, yet generic, framework is presented and supported with the POSTEDFIT capability elements. The paper is concluded with future research recommendations in Section 7.

2. Definitions and context

The arena of the cyberspace is quite diverse with different interpretations, capabilities, and skills requirements. The understanding of this operational environments has an influence on how African nations could build capabilities for securing their tailored militaries.

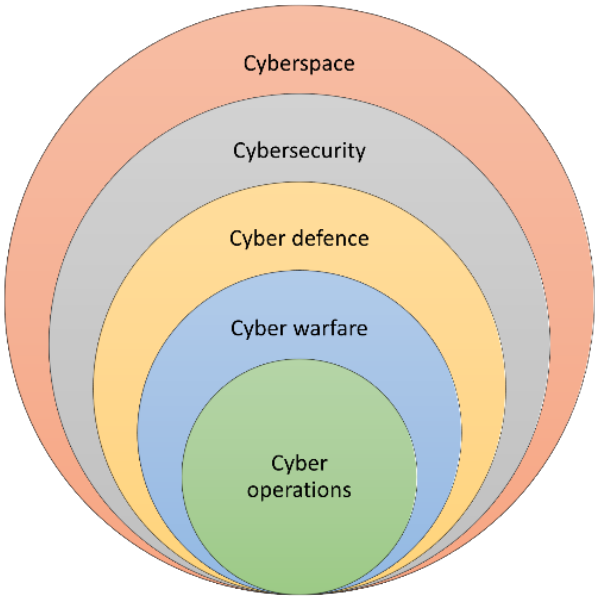


Figure 1: Cyber Warfare Context (source: authors)

In essence as depicted in Figure 1, cyberspace is a domain of operation and is simply defined as a *"notional environment that enables people and systems to communicate over interconnected networks"*, according to a Dictionary of Psychology (Colman, 2014). In the military context, it is classified as the fifth operational domain of warfare (Hall, 2016) and interacts with other domains of warfare such as space, air, land, and sea.

Within the cyberspace, a plethora of capabilities could be observed from a security-perspective, and the most common capability is that of cybersecurity that deals with solutions that minimizes danger or threat to organizational assets from threat actors. This capability is a necessity across all economic sectors as well as in the military. In terms of this paper, two positions are considered when it comes to offensive cyber, and that is cyber defence and cyber warfare**.** In this research, we perceive the cyber defence capability to be adopting in the main a defensive posture within the cyber battlespace as relied upon by the North Atlantic Treaty Organization (NATO, 2022).

On the contrary, cyber warfare, which is the focus on this paper, is define as a capability of a nation-state to exploit another nation-state's communication networks, computer systems and/or other critical information infrastructure for purposes of causing delay, disruption, and/or damage (Wang, 2023). According to (Oosthuizen & Roodt, 2008), capability may be conceived of as comprising nine POSTEDFIT (Personnel, Organisation, Sustainment, Training, Equipment, Doctrine (Policies), Facilities, Information and Technology) constituent elements. Lastly, cyber operations are critical in every phase of the modern warfare and are in this paper loosely defined as non-lethal elements of warfighting functions in the cyberspace that produce specific effect on a target, such as deny, disrupt, and destroy (Shankar, 2023).

The understanding of cyber warfare is therefore related to the understanding of the environment within which this war takes place, the cyberspace. The United States of America (USA) Department of Defence (DOD) through its Joint Chiefs of Staff, in the Joint Publication 3-12, for cyber operations, highlights that cyberspace, while part of the information environment, is dependent on the physical domains of air, land, maritime, and space. This publication further indicates that the cyber operations use links and nodes located in the physical domains and perform logical functions to create effects first in cyberspace and then, as needed, in the physical domains (US DOD, 2018).

3. Importance of cyber warfare capabilities in modern conflicts

The ongoing Russia-Ukrainian conflict and growing geo-political tensions places a new emphasis on critical industries and national security, leading to more strict security requirements and restrictions. According to Burt (2022), an increase in military-coordinated cyber-attacks, and a continued growth of conflict in cyberspace has been observed over the recent past. In the Russia-Ukraine conflict that escalated in early 2022, we have observed that this conflict is hybrid using physical weapons as well as cyber-attacks. For instance, in January 2022, it is reported that a cyber-attack targeted the Ukrainian government, hitting 90 websites and deploying

malicious software masquerading as ransomware to damage dozens of computers in government agencies. Moreover, in February 2022, a DDoS attack knocked down websites belonging to the Ukrainian Defence Ministry and two of the country's largest banks offline. In China and US, we have seen threat actors spreading mobile malware to citizens using fake cell phone towers, as well as sending hate-SMS-messages from rogue base stations (Chirgwin, 2017; Koebler, 2015).

As per the recent research study by Voo, Hemani and Cassidy (2022), the National Cyber Power Index (NCPI) demonstrates that nation-states are serious about building cyber warfare capabilities for the modern conflicts. The NCPI measures nations demonstrated and potential capability in strategies, defensive and destructive operations, resource allocation, private sector capabilities within a country, workforce, and innovation. Possible cyber operations that a nation would be having to demonstrate cyber power include surveillance and monitoring of domestic groups, strengthening and enhancing national cyber defences, controlling and manipulating the information environment, foreign intelligence collection for national security, growing national cyber and commercial technology competence, destroying or disabling an adversary infrastructure and capabilities, defining international cyber norms and technical standards, and amassing through cyber operations.

Figure 2 below shows the scatter plot of cyber power rankings of 30 countries, and it is evident that African nation states are non-existent, except for Egypt, who is shown as having a lower capability and lower intent in cyber power, whilst countries such as the United States and China are well advanced having higher capability and higher intent to use cyber means to demonstrate power in the cyberspace. These countries have a direct interest in cyber warfare capabilities for Africa as observed by their support and/or attacks in the African cyber space.

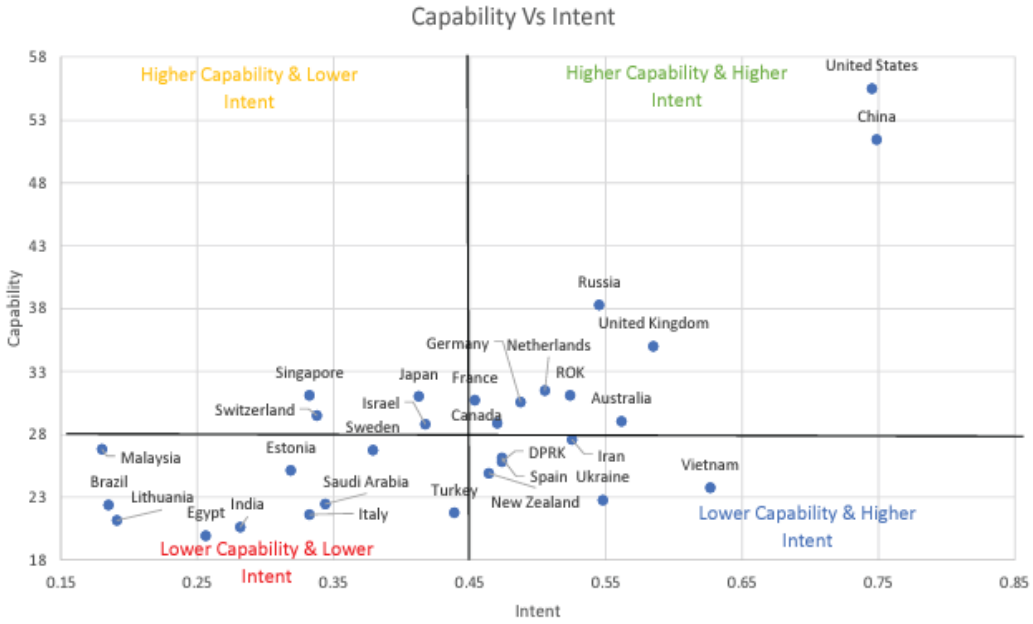


Figure 2: Nations Cyber Warfare Capability vs Intent (Voo et al, 2022)

3.1 Overview of existing cyber warfare capabilities in Africa

In Africa, we have started to observe the establishment of cyber warfare capabilities and use of cyber operations by extremists' groups and nation-states. For example, a Yemen Cyber Army claimed responsibilities for cyber-attacks on several websites from various government entities and ministry of defence in Mozambique in Mozambique (AllAfrica, 2022).

In Nigeria, it is reported that Boko Haram hacked the personnel records database of Nigeria's secret service (Baken, 2013) and this has subsequently led to the Nigerian Chief of the Army (Nigerian Army, 2022) ordering the creation of the Cyber Warfare Command as well as the Cyber Warfare School.

In 2020, Samme-Nlar (2020) reported that both state actors and non-state actors are increasingly targeting African states using cyber weapons. An example of the attack on the Ethiopian cyberinfrastructure by the

Egyptian non-state actors over a grand renaissance dam dispute was provided. In this case study, it is also made clear that African states are unprepared to deal with cyber-attacks.

In South Africa, the last Defence Review was done in 2014 (South African Government, 2014), and when it comes to cyber warfare, the review suggested that South Africa requires the protection of its cyber-domain, through a comprehensive information warfare capability, integrated into its intelligence-related information systems at the international, national and defence levels. In line with the Defence Review, the National Policy Cybersecurity Policy Framework (NCPF) was published in 2015 (State Security Agency, 2015), and it defined DOD's overall responsibility as the coordination, accountability, and implementation of cyber defence measures in South Africa such as the establishment of the Cyber Command. By 2022, it was apparent based on media reports and parliamentary feedback that the cyber command still faces several challenges, such as lack of resources and skills (DefenceWeb, 2023).

In South Africa, researchers further report that cyber-attacks against state and private organizations have increased by over 95% between 2010-2020 with attacks ranging from website hacks, denial of service attacks, data breaches, ransomware attacks (Pieterse, 2021). Researchers also report that South African public platforms such as websites have seen an increased targeted cyber-attacks mainly because over 80% of public websites are vulnerable (Mtsweni, 2015). As such, governments across the globe continue to invest in cyber warfare capabilities. Over 60 countries, a number that is rising, currently have some mechanism to play in the digital warfare and intelligence gatherings. Nevertheless, African countries always lag in this regard.

4. Research Methodology

The research presented in this paper follows the Design Science Research (Hevner et al., 2010) which subscribes to the concept of an artefact. In this paper, the comprehensive framework for developing cyber warfare capabilities is considered as an artefact. The artefacts are derived through reference to the literature, existing frameworks, as well our experience in the cyber security and military environment within the African continent.

Design science research methodology was chosen for this research work because it is a "*problem-solving approach that seeks to enhance human knowledge via creation of innovative artefacts*" (vom Brocke; Hevner; Maedche, 2020). The key steps that were followed for the purposes of this research study will focus on people, processes, and technology.

These key steps will be enhanced by a POSTEDFIT approach that is commonly used in building military capabilities (Willers et al, 2011; Mtsweni et al, 2018). In the design phase of the DSR, a cyber warfare capability development framework is proposed and evaluated through use case scenario analysis as per the guidelines found in (Hevner et al., 2010).

It is important to note that not all POSTEDFIT elements are the same under different contexts. It is therefore critical to look at them closely under each environment, and they may also not be weighted the same by different nation states depending on their intentions and missions in the cyber space. It is worth noting that in this paper, leadership and budget elements are in the people and support elements respectively.

4.1 Systems thinking approach.

Systems thinking approach (Litster, Hurst, & Cardoso, 2023) is applied in this research to understand and develop a comprehensive framework for developing cyber warfare capabilities, which are complex in nature. The dynamic nature of the cyber threats, and the evolving cyberspace continues to challenge the traditional ways of determining solutions and thus a systems thinking approach has been chosen a guideline to evolve the comprehensive cyber warfare capability framework.

This approach suggests that we need to understand the functionality required from the strategic direction provided by the organisation before we think about the solutions required. The ability of any military and/nation, to conduct operations in the cyberspace must be guided by that nation's national security strategy. From this understanding, it also follows that the cyber warfare capabilities are derived from the functionality implied in the missions defined to achieve the set national strategic objective (Smit et al, 2012). Figure 3 shows an example of the decomposition of the national objectives to capabilities. Military capabilities are then decomposed further into system of systems, then systems, right down to components. This approach is therefore demonstrated in the proposed framework.

Figure 3: Systems thinking for cyber warfare capability development (Smit et al, 2012).

5. Cyber Warfare Capability Mapping

In this section, we highlight and briefly discuss the different approaches that are used in cyber operations, and we map these against different cyber warfare capability objectives and functions.

Cyber warfare capabilities may differ from one nation state to the other as observed in the National Cyber Power Index (Voo et al, 2022) depending on the capability levels and intentions. The overall NCPI assessment measures the 'comprehensiveness' of a country as a cyber actor. Comprehensiveness, in the context of NCPI, refers to a country's use of cyber to achieve multiple objectives as opposed to a few. The most comprehensive cyber power is the country that has (1) the intent to pursue multiple national objectives using cyber means and (2) the capabilities to achieve those objective(s) (Voo et al, 2022).

In addition, the cyber domain has seen a myriad of “kill chains” that demonstrate different tactics that could make a nation-state successful or not when they are conducting cyber operations.

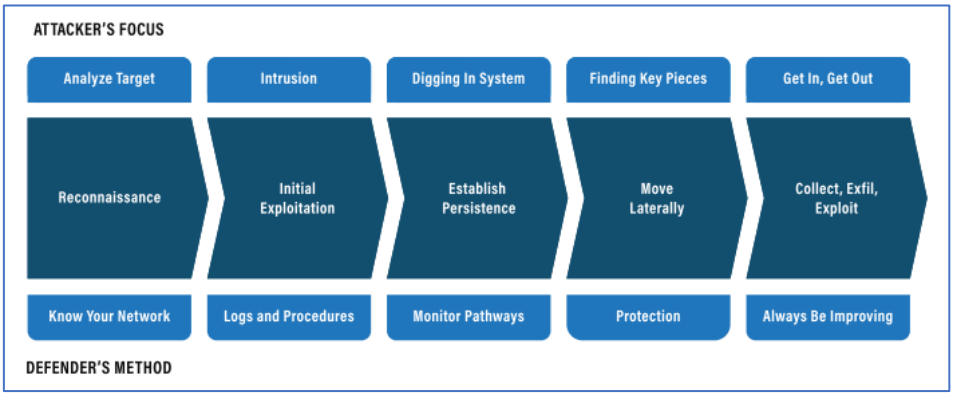


Figure 4: The Intrusion Model (Tarnowski, 2017)

As depicted in Figure 4, it is also observed that cyber operations share similar characteristics with conflicts in the physical domains but differs in others. For instance, information and communication technologies have dual use where they may be exploited to aid common business operations but could also be exploited to advance strategic military missions. Whilst in the physical realm, miliary equipment (e.g., fighter aircraft) has a strategic military purpose in mind. In the cyber domain, the universe of adversaries is also quite wide, unlike in the physical domain where location, proximity or historical conflicts may provide signs of who is likely to attack you or not. The essence in the intrusion model is that it is two sides of the same coin, and this means that nation states need to have both the defensive and offensive posture. As an attacker, the military, in the cyberspace, needs to be able to analyse the target through intelligence gathering, but at the same time the same nation needs to fully understand its assets (e.g., networks or critical information infrastructure) including their whereabouts and status at all time or failure to know this can result into devastating defeat in the cyberspace.

In the cyber space, militaries need to deal with myriad of adversaries and not only nation state actors. These may include hackers, terrorists, businesses, social groups, criminals, and even unsuspecting computer users. The inability to define the enemy is the reason some of the nations may deal with the consequences of cyber-attacks rather that then root causes. As narrated by Sun Tzu in the Art of War: *“if you know the enemy and know yourself,*

you need not fear the result of a hundred battles. If you know yourself but not the enemy, for every victory gained you will also suffer a defeat. If you know neither the enemy nor yourself, you will succumb in every battle." Thus, it is critical that when African militaries are building cyber warfare capabilities, they do not only focus on only the nation-states as an adversary, but other threat actors as well that could be nation-state sponsored such as cyber hacktivists groups.

Lastly, in the cyberspace, nations' assets may be infiltrated in peace time and used in war time, thus the situational awareness in the cyberspace needs to be wider than just an internal view. This also means it is becoming more difficult to have a complete view on who may attack a nation or already attacked a nation in the cyberspace. It is also complex to attribute responsibility of cyber-attacks, as in the cyberspace attackers can even use various techniques including proxies to divert attention. Moreover, the motives for cyber operations can be quite wide including economic, military, national secrets and political advantage, and the cyberspace allows for cyber operations to be ensued based on any of these motives, because in the cyberspace, offensive operations are significantly easier than defence, mainly because anyone can advance them, whilst in the traditional warfare, defence is the default position, and a successful attack requires supremacy in people, technology, doctrine, or strategy.

6. Comprehensive framework for developing cyber warfare capabilities for African Militaries

6.1 Cyber warfare capability framework requirements for operationalization

The framework for cyberwarfare capability, and the military's role in the cyberspace, follows the same analogy for war in other domains. The framework to be considered must cover the full spectrum of the ability of the military to conduct operations in the cyberspace. These operations must address both the offensive and defensive capabilities.

The ability by the military to conduct operations in the cyberspace, requires the following (this understanding is on the premise that higher political and strategic intent is understood including the military in the cyberspace):

- The understanding of the political, and strategic objectives of the country.
- The understanding of the cyberspace, what constitutes the domain.
- The legal framework, governing roles of various entities.
- The ability to secure its own systems.
- The ability to conduct reconnaissance and gather intelligence against adversaries.
- The ability to do target acquisition.
- The ability to launch offensive actions against target adversaries.
- The ability to defend offensive actions of the adversary or its associates.
- Recovery from an onslaught or cyber-attacks.
- Withdrawal from cyber operations when the situation dictates.
- Continuous improvement of the capability.

Over and above these requirements, the availability of resources such as a budget for the military is critical to establish a cyber warfare capability. In has been noted, for instance, in South Africa that Defence Intelligence has struggled to setup a fully functional Cyber Command due to lack of budget allocated for such a capability (DefenceWeb, 2023).

6.2 Design of the framework

In designing the comprehensive framework for developing cyber warfare capabilities for the African militaries, several considerations were made, including existing capabilities based on publicly available information, capability objectives in the cyberspace, ranking of African militaries in relation to the National Cyber Power Index, and national and international norms and standards.

In formulating this framework, the authors followed a three-layered approach that focused *on 1) national strategic goal in the cyberspace, 2) cyber warfare capability objectives for a nation-state, and 3) minimum cyber warfare capability functions* that nation-state need to possess to demonstrate a cyber warfare capability in the battlespace. Additionally, the capability functions were mapped against the POSTEDFIT capability elements as depicted in Figure 5.

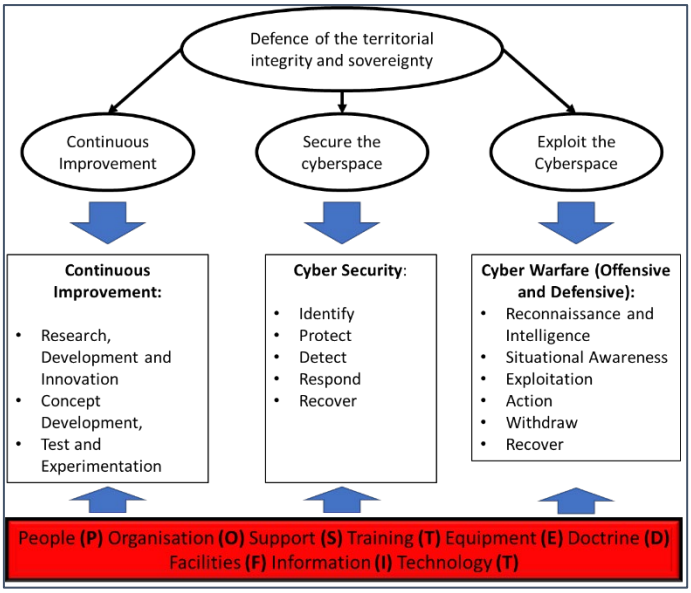


Figure 5: Proposed comprehensive cyber warfare capability framework.

In brief, the model suggests that for African militaries to establish, deploy and sustain their cyber warfare capabilities, the main goals in the cyberspace should be about defending territorial integrity and sovereignty. These goals are no different to militaries in defending their nations in the air, sea, land, and/or space. In particular, the core objectives for playing an effective role in the cyberspace, African militaries need to ensure that there are continuous improvements, through RD&I, concept development and experimentation in building, executing, and sustaining the cyber warfare capability. This capability is seen in two lenses: *(1) securing the cyber space (i.e., taking a security and protection approach)*, and *(2) exploiting the cyberspace to gain territorial and sovereignty advantage through offensive means*.

Any nation-state that needs to exert its power in the cyberspace must have the capabilities as depicted in the cyber kill chain or intrusion model (see Figure 4). The preliminary function is reconnaissance and threat intelligence. African militaries need to understand the enemy through continuous scouting of intelligence and areas of interest. This intelligence would allow for situational awareness across all domains of war and enable the forces to focus on the crown jewels of the enemy in the cyberspace. The exploitation of the enemy needs to be executed both in peace and war time, but in a covert manner. Countries such as China, Russia and United States are super-powers in the cyberspace, and they also take advantage of the capable privately accessible offensive capability to exploit their adversaries in peace times and easily destroy the target in war times. And as such, African militaries need to also strategic form partnership with local and international industries to enhance and continuously improve the cyber warfare capability.

6.3 Understanding the cyber warfare capability development framework

To implement the framework, we draw up a capability matrix that is mapped to capability attributes (functions) and capability elements. These elements are weighted, because we are of the view that they are not of equal importance in the cyber domain, and in certain instances their importance is given effect by the cyber mission. In this research paper, we recommend that any capability should have a higher value or proportion on people, technology, and processes. Using the POSTEDFIT framework, this perspective would map to people, doctrine, and technology.

6.4 Framework use case scenario analysis

To demonstrate the utility of the framework, we choose *Country X in Africa* that is still establishing its cyber warfare capability and evaluated it using the proposed framework. It should be noted that we demonstrate the framework by using officially known capabilities and in instances where no information is available assumptions using lessons from cyber-attacks that we have observed online.

Table 1: Framework maturity level index for POSTEDFIT elements and overall capability

	POSTEDFIT Elements		Overall Capability Maturity	Range	Scale
			MLI 0	0%-20%	Very Weak
3	Full		MLI 1	21%-40%	Weak
2	Intermediate		MLI 2	41%-60%	Moderate
1	Limited		MLI 3	61%-80%	Strong
0	None		MLI 4	81%-100%	Very Strong

In demonstrating the framework, the first step is measuring the cyber warfare capability using the capability attributes scales chosen for each of the elements under each capability as depicted in Figure 5. The framework adopts a 3-level ranking system for the POSTEDFIT capability elements and associated functions, and these are classified as *0) No functions implemented 1) Limited functions 2) Intermediate functions and 3) full functions* (see **Table 1:** *left*). This means that to assess if a country has cyber security or cyber warfare capability, one will rate the capability associated functions between 0-3. These are then aggregated to determine the maturity level of each sub-capability. In addition, the framework uses the 5-level ranking system as found in the NIST cybersecurity maturity model (Almuhammadi, 2017), classified as maturity level indicator (MLI) (MLI0-MLI4) *(*see Table 1: *right*).

In the proposed framework, the authors opted to use a "four force model[1]" scales to narratively indicate the meaning of the different MLI levels, and these are very weak for none to minimal capability to very strong for advance or full capability. For example, within a range of 0-20, this means that the maturity level of the cyber warfare capability for Country X is *Very Weak*, whilst 61-80 will be considered as S*trong*. As already indicated, these ranges could be adjusted to suit the dynamics of different African states. It also needs to be borne in mind that this framework could be used to measure the "AS-IS" cyber warfare capability or alternatively the "TO-BE" capability of a country of interest.

Table 2: Implications and implementation of the framework

				Capability Elements									
				People	Organisation	Support	Training	Equipment	Doctrine (Policies and Procedures)	Facilities	Information	Technologies	Maturity
Capability Maturity Level Indicator	Goals	Capability	Capability Attributes	15%	10%	10%	15%	5%	15%	5%	10%	15%	100%
	Security	**Ability to secure the cyber space**											45%
		30%	Identify	2	2	2	2	2	2	2	2	2	
			Protect	2	1	1	2	1	2	2	2	2	
			Detect	1	1	1	2	1	1	1	2	2	
			Respond	1	1	1	1	0	1	0	2	1	
			Recover	1	0	0	1	0	1	0	2	0	
	Exploitation	**Ability to conduct Offensive and Defensive Cyber Operations**											34%
39%		50%	Reconaissance and Intelligence	2	2	2	2	1	2	2	2	2	
			Situational Awareness	1	0	1	1	1	2	1	2	1	
			Exploitation	1	0	1	1	1	0	0	2	1	
			Cause of Effects	1	1	1	1	1	1	0	2	1	
			Sustain and Defend	1	1	0	1	0	1	0	1	1	
			Withdrawal	1	0	0	0	0	0	0	1	1	
	Continuous Improvement	**Ability to keep abreast with the evolving cyber battlespace**											39%
			Research, Development and Innovation	2	1	1	2	2	1	2	2	2	
		20%	Concept Development	2	1	1	0	1	0	1	2	1	
			Test and Experimentation	1	1	1	1	1	0	1	2	1	

The application of the model as demonstrated in Table 2 above is progressive, and accumulates from the scores for the capability elements, adding up to overall cyber warfare. As indicated in the Table 2, the total scores for the POSTEDFIT elements per capability, are weighted and thereafter added to give a total score for the individual capability attributes. For example, in Table 2, it can be observed that Country X scores *Moderate* (45%) on the

[1] https://revisionscience.com/a2-level-level-revision/physics-level-revision/particles-radiation-quantum-phenomena/four-force-model

sub-capability to secure the cyber space, but is measured to be *Weak* (34%) on the ability to conduct offensive cyber operations as well as in their ability to keep abreast with evolving cyber battlespace.

After the individual sub-capabilities are scored, their scores are weighted against the determined proportion of contribution of each capability to the overall capability. In this paper, the weighting for sub-capability one is 30%, sub-capability 2 is 50%, and sub-capability 3 is 20%. The outcome is the overall capability maturity level indicator, which in this example in Table 2 is 39%. This score indicates the overall level of a nations' ability to defend the territorial integrity, and sovereignty of the country within cyberspace. In this instance, *Country X* is measured to be *Weak* on the overall capability, and this implies that they may not be able to effectively operate offensively and/or defensively in the cyber battlespace.

7. Conclusion and Future Research

The evolution of the operational environment, and the emergence of the cyberspace as a domain of war, remains critical to the strategic direction of many militaries. The African battlespace is not immune to this development, and the global nature of the cyberspace forces all African militaries to include cyberwarfare capabilities in their portfolio. This paper provided a framework, to define these capabilities, in terms of the capability attributes, and the capability elements, and provides a link between the capabilities and the national strategic objectives. The framework is also instrumental in measuring the level of maturity of the capability, and determines capability gaps, which will help nations draw up development plans to address the gaps. The framework is at this stage still simplistic, in terms of the calculations and the definitions of various measures used in the calculations. This implies that the validation of the model will still be done as part of the continuous research on the model. Further to the proposed framework, more research must still be undertaken to understand the cyber warfare capabilities, and related attributes, to enable for measuring effectiveness and/or efficiency of the capabilities.

References

Ajijola, A.H and Allen, N.D.F. 2022. African Lessons in Cyber Strategy. African Centre for Strategic Studies. Available online: https://africacenter.org/spotlight/african-lessons-in-cyber-strategy/

AllAfrica. 2022. Mozambique: Hackers Demand Ransom. Available online: https://allafrica.com/stories/202202230552.html

Almuhammadi, S. and Alsaleh, M., 2017. Information security maturity model for NIST cyber security framework. *Computer Science & Information Technology (CS & IT)*, *7*(3), pp.51-62.

Baken, D.N. 2013. Cyber warfare and Nigeria's vulnerability. E-International Relations. ISSN 2053-8686

Burt, T.2023. Nation-state cyberattacks become more brazen as authoritarian leaders ramp up aggression. Microsoft, Nov 4. 2022. Available online: https://blogs.microsoft.com/on-the-issues/2022/11/04/microsoft-digital-defense-report-2022-ukraine/

Chirgwin, R. 2017. Fake mobile base stations spreading malware in China. The Register. Available online: https://www.theregister.com/2017/03/23/fake_base_stations_spreading_malware_in_china/

Colman, A.M. 2014. A Dictionary of Psychology (3 ed.) Oxford University Press. ISBN: 9780199534067

DefenceWeb. 2023. SANDF Cyber Command operating in limited space. Available online: https://www.defenceweb.co.za/featured/sandf-cyber-command-operating-in-limited-space/

Hall, S., 2016. *Cyberspace at the operational level: warfighting in all five domains*. Joint Military Operations Department Newport United States.

Hevner, A., Chatterjee, S., Hevner, A. and Chatterjee, S., 2010. Design science research in information systems. *Design research in information systems: theory and practice*, pp.9-22.
https://www.e-ir.info/pdf/43935

Koebler, J. 2015. "The FBI admits it uses fake cell phone towers to track you." Motherboard Tech by Vice. Available from https://www.vice.com/en/article/jp5azg/fbi-admits-it-uses-fake-cell-phone-towers-to-track-you

Litster, G., Hurst, A. and Cardoso, C., 2023. A Systems Thinking Inspired Approach to Understanding Design Activity. In *Design Computing and Cognition'22* (pp. 161-177). Cham: Springer International Publishing.

Mtsweni, J., Gcaza, N. and Thaba, M., 2018, September. A unified cybersecurity framework for complex environments. In Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists (pp. 1-9).

Mtsweni, J., 2015. Analysing the security posture of South African websites. In *2015 Information Security for South Africa (ISSA)* (pp. 1-8). IEEE.

NATO. 2022. Cyber Defence. Available Online: https://www.nato.int/cps/en/natohq/topics_78170.htm

Ndebele, l. 2023. US Army to train some African countries in cyber, land and maritime defence. Available from: https://www.news24.com/news24/africa/news/us-army-to-train-some-african-countries-in-cyber-land-and-maritime-defence-20230116. Last Accessed: 01 February 2023

Nigerian Army. 2022. Cyber Warfare Command. Available Online: https://army.mil.ng/?page_id=5714

NIST, 2018. Framework for improving critical infrastructure cybersecurity, version 1.1. Available online: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf

Oosthuizen, R., & Roodt, J. 2008. Credible Defence Capability: Command and Control at the Core. Land Warfare Conference

Pieterse, H. 2021. The cyber threat landscape in South Africa: A 10-year review. *The African Journal of Information and Communication (AJIC), 28*, 1-21. https://doi.org/10.23962/10539/32213

Samme-Nlar, T. 2020. The future of armed conflict in Africa: what cyber-attacks on Ethiopian government tells us. Available online : https://aanoip.org/the-future-of-armed-conflict-in-africa-what-cyber-attacks-on-ethiopian-government-tell-us/

Shankar, A. 2023. Offensive Cyberspace Operations: using artificial intelligence and kill chains to analyze the effects of MAGTF execution authority. Marine Corps Gazette. February 2023. Available from: https://www.hoover.org/sites/default/files/research/docs/Offensive%20Cyberspace%20Operations%20-%20MCG%20-%20Feb%202023%5B93%5D.pdf

Smith. C and Oosthuizen R, 2012. Applying systems engineering principles towards developing defence capabilities. Available Online: https://incose.onlinelibrary.wiley.com/doi/abs/10.1002/j.2334-5837.2012.tb01383.x?casa_token=W7MQrWozFTAAAAAA%3ANtijJM3-_Wkq1M997uQg4_VEgcKJLZX_tQoZHkJQKDtGL8gVP57gcjvDqKBwQ2nf3XTXrl0F3iuNwCM

South African Government. 2014. South African Defence Review 2014. Available online: https://www.gov.za/sites/default/files/gcis_document/201409/dfencereview2014.pdf

State Security Agency. 2015. The National Cybersecurity Policy Framework (NCPF) for South Africa. Available online: https://www.gov.za/sites/default/files/gcis_document/201512/39475gon609.pdf

Tarnowski, I., 2017. How to use cyber kill chain model to build cybersecurity? European Journal of Higher Education IT.

US DOD. 2018. Joint Publication 3-12: Cyber Operations. Available from: https://publicintelligence.net/jcs-cyberspace-operations/

vom Brocke, J., Hevner, A. and Maedche, A., 2020. Introduction to design science research. In *Design science research. Cases* (pp. 1-13). Springer, Cham.

Voo, J; Hemani, I; and Cassidy, D. 2022. National Cyber Power Index 2022. Belfer Center for Science and International Affairs. Harvard Kennedy School. [Online] Available from: https://www.belfercenter.org/publication/national-cyber-power-index-2022, Last Accessed: 01 February 2023

Wang, A. 2023. Cyberwarfare: the final frontier of conflict. Havard Model Congress, Boston, USA. Available from: https://www.harvardmodelcongress.org/s/HMC-2023-Cyberwarfare-dhtw.pdf Last Accessed 01 February 2023

Willers, C. J., Al-Ghamdi, A. A, Bezuidenhout, D. F., and Al-Hosiny, N. M. 2011. "Establishing an infrared measurement and modelling capability," *2011 Saudi International Electronics, Communications and Photonics Conference (SIECPC)*, Riyadh, Saudi Arabia, 2011, pp. 1-6, doi: 10.1109/SIECPC.2011.5876978.

Cybersecurity Through Thesis in Laurea University of Applied Sciences

Ilona Frisk [1], Harri Ruoslahti [1] and Ilkka Tikanmäki [1, 2]
[1]*Security and Risk Management, Laurea University of Applied Sciences, Espoo, Finland*
[2]Department of Warfare, National Defence University, Helsinki, Finland

Ilona.frisk@laurea.fi
Harri.ruoslahti@laurea.fi
Ilkka.tikanmaki@laurea.fi

Abstract: Information technology and its applications surround us and those have become crucial to our lives. However, the understanding of the digital world is not as strong. Successful and functional cybersecurity is a vital component for the defence of a civilised society. This study looks at how cybersecurity has been handled in thesis written at one University of Applied Sciences and what kind of topics have been chosen by thesis writers, and what is written about cybersecurity in them to understand how cybersecurity is seen in higher education. The goal of this paper was to find out how cybersecurity has been handled in theses and what kind of topics have been chosen by thesis writers. The two research questions are: what theses have been published that handle cybersecurity; and how does cybersecurity in them? As typical of a case study, attention is paid to a small number of cases (n = 15) attempting to describe the phenomenon they represent. Of the fifteen theses, two were master's and thirteen bachelor's theses, and mostly completed in Safety, Security and Risk Management, Security Management, and Business information technology programmes. Based on the results in this case, cyber security is being examined or developed from several, different points of view and in multidisciplinary ways.

Keywords: Cybersecurity, comprehensive security, cybercrime, resilience

1. Introduction

Information technology and its applications surround us in today's world. These have become very important for our work and civilian lives, and we have become increasingly dependent on them (Shoemaker and Conklin, 2011). However, this does not mean that our understanding of the digital world is strong, but quite the opposite there is a recognized gap in understanding what cybersecurity is and what affects it, despite efforts to increase overall security. Functional and successful cybersecurity is a vital component in the defence of civil society and its organizations, and each individual is part of a community and of society, and part of this defence which is why it is important to understand what cybersecurity is (Bowden, 2010, p. 16; Burgess, 2010; Limnéll, Majewski and Salminen, 2014).

Project Dynamic Resilience Assessment Method including a combined Business Continuity Management and Cyber Threat Intelligence solution for Critical Sectors (DYNAMO) builds on the work of European Network of Cybersecurity Centres and Competence Hub for Innovation and Operations (ECHO) to understand the role of e-skills in society within the societal impacts of project outcomes and cyber security.

[1] https://orcid.org/ 0000-0001-9726-7956
[2] https://orcid.org/0000- 0001-8950-5221

To understand how cybersecurity is visible in the context of higher education, as seen by soon to graduate future work life experts this study looks at how cybersecurity has been investigated in theses written at Laurea University of Applied Sciences (UAS), and what topics the authors of these theses have chosen. Laurea graduated some 800 bachelor's and master's students in 2021, from 43 degree programs fields ranging from social services and nursing to business and administration, security and risk management, and business information technology, which are most likely to choose cyber security topics (Laurea University of Applied Sciences, 2022). All Laurea graduates can select an interesting and relevant topic of study or development on which they are, as part of their degree, publish a thesis report. This study examines the theses published in 2021 and beginning of 2022 to understand how cybersecurity is seen in higher education, by looking at how they write about cybersecurity.

The research questions (RQ) of this study are:

RQ1: What published theses are on cybersecurity?

RQ2: How does cybersecurity appear in them?

2. Cybersecurity

The importance of the cyberworld makes cybersecurity a topical issue it is, however, not clear to everyone what cybersecurity stands for. Based on the literature examined, it appears that little research has been done on what topics are studied at higher education institutions (HEI). This section provides an overview on the literature on cybersecurity.

Several studies in scholarly literature find that cybersecurity is multidisciplinary and containing concepts from several fields (Craigen, Diakun-Thibault and Purse, 2014) e.g., business management, behavioural studies, and technical and software engineering (Shoemaker and Conklin, 2011). The first part *cyber* is used as a prefix refering to the digital world of bytes (Limnéll, Majewski and Salminen, 2014). It originates from the word cybernetics, which is the study of computers and automation systems communications, and simultaneously used with the concept cyberspace, a virtual and intangible space that exists by merging a variety of electronic communication systems and networks in an electronic space that is not part of the known and seen physical space; cyberspace is however, bound by physical elements and cannot exist without wires, cables or electromagnetic waves (Cavelty, 2010).

The dictionary defines the second part *security* as being safe and protected from danger and threats or measures used to accomplish that state (Lexico, 2022) One definition for cybersecurity is provided by (Craigen, Diakun-Thibault and Purse, 2014, p. 17) as "organization and collection of resources, processes, and structures used to protect cyberspace and cyberspace-enabled systems from occurrences that misalign property rights". Literature recognizes multiple definitions of cybersecurity, the definitions by e.g., (Cybersecurity and Infrastructure Security Agency (CISA), 2019), (Kaspersky, 2022), (National Cyber Security Centre, 2022) all look at cybersecurity being the protection of computers, devices, and services.

Cybersecurity is needed to make the virtual world secure and safe; cybersecurity can be seen as the use of technology and legislation to protect and manage information (Ruoslahti and Tikanmäki, 2022); or processes and measures to protect cyberspace and its systems and other physical aspects, devices and software from possible threats (Craigen, Diakun-Thibault and Purse, 2014). Thera are similarities between information security and cybersecurity as both of their measures have three goals: to protect the confidentiality, integrity, and availability of information (Cavelty, 2010). It is not possible to separate between the physical and the virtual or digital elements of cybersecurity, as cyber events can have physical consequences and tangible effects even though cyberspace in itself is intangible (Shoemaker and Conklin, 2011; Limnéll, Majewski and Salminen, 2014). Cybersecurity and security in general should be embedded as a solid part of all organizational processes so that security dimensions are considered in all actions and operations from the very beginning, not only when a new digital service is otherwise ready to be implemented (Limnéll, Majewski and Salminen, 2014).

Cybersecurity needs to be consistent and continuous process because cyber threats and vulnerabilities are constantly new and emerging (Cavelty, 2010). Threats will continue to exist, and therefore cybersecurity measures are forever needed, as the goal of cybersecurity is to protect applications and cyberspace from various threats that could compromise its safety (Craigen, Diakun-Thibault and Purse, 2014). Besides a technical issue, cybersecurity is mostly a strategic and political, with situational awareness, defined direction and guidelines in both organizations and society; making cybersecurity part of comprehensive security that belongs to all of us (Limnéll, Majewski and Salminen, 2014). On a larger scale cybersecurity is close to national security, the differences are in the actors, scope, and funding, which both aim to create conditions that are free from either imagined or existing danger (Cavelty, 2010).

One internationally recognized threat to cybersecurity is cybercrime (Mohammed, 2015). The number of crimes and costs caused by cybercrime have been continuously increasing since their first appearance since personal computers became popular in the 1970's and 1980's. In the early stages, cybercrimes were divided in those threatening businesses and those affecting in national security (Cavelty, 2010), but modern definitions of cybercrime involve all illegal and criminal acts against computer data, systems, as well as unauthorized access, modification or impairment of digital or computer systems, which do not have geographical boundaries and only leave digital traces (Mohammed, 2015; Payne, 2020).

Besides its many risks and threats, the cyberworld provides organizations and businesses many new opportunities to grow and add value, so decision makers especially need updated knowledge and up-to-date situational information to build stability between threats and opportunities within this cyberspace all-around us (Limnéll, Majewski and Salminen, 2014). Information and communications technology (ICT) is important in building organizational competitiveness (Mihalic and Buhalis, 2013) by promoting productivity and

competitiveness with the generation of knowledge, processing information, and organizational learning (Hortovanyi and Ferincz, 2015). Knowledge transfers can become promoted through ICT skills of the members of the organization (Salleh *et al.*, 2012), because knowledge flows and organizational learning can be enhanced with ICT (Škerlavaj, Dimovski and Desouza, 2010; Zhao and Kemp, 2013).

ICT skills can upgraded to the use of the various Knowledge Management (KM) technologies and ICT tools needed to transfer and share information within and outside the organization by proper ICT trainings (Conkova, 2013; Isidro-Filho *et al.*, 2013). Building skills and competences are constructive processes that use and recognize previously adopted competences of learners, and aim at helping them better navigate the cyber domain (Aaltola and Taitto, 2019). When having appropriate ICT knowledge, workers can capture, store and share organizational knowledge that makes their expertise better available to the organization, its network and the surrounding society (Im, Porumbescu and Lee, 2013), and organizations can develop the skills needed to absorb state-of-the-art knowledge from external sources (Cupiał *et al.*, 2018).

3. Methodology

This paper presents a multiple case study. Case studies usually concentrate on one or some cases that are related to each other. This study considered cybersecurity themed theses published at Laurea University of Applied Sciences as cases (Hirsjärvi *et al.*, 2000) to understand how cybersecurity has been investigated in them, and what topics have been chosen by their authors. Laurea students may select an interesting topic of study or development that has work life relevance. This can be done based on their knowledge, taught courses, proposed by teachers, Laurea research and development projects, authority organizations, or private companies.

The first stage of purposeful sampling used in this case study (Merriam and Tisdell, 2015) was to search theses from Theseus, the common database of Finnish universities of applied sciences, where abstracts, bibliographic information, and full texts of the theses are publicly available. The search included theses written at Laurea during a set period of time in Finnish and English. The theses were searched from the database with search terms *cybersecurity* and its Finnish translation *kyberturvallisuus* appearing in either the main title, abstract or as a keyword.

Thesis between 1.1.2021 and the day of the search 21.4.2022 were included, with an initial result of 52 theses that were organized according to their publication dates. Next, inclusion criteria that at least one search term or part of it be included in either the main title, abstract or keywords, were applied rendering 37 thesis reports that did not meet the criteria. Thus, fifteen thesis reports qualified for the final sample.

The qualitative analysis of the collected data included reading their full texts to examine how cybersecurity is investigated in each of these 15 theses. Common themes or categories that could be recognized were identified, as finding common themes or categories and building larger entities out of them to be able to describe the phenomenon of interest is typical to qualitative data-analysis (Merriam and Tisdell, 2015), which was deemed a suitable method to find answers to the research questions. The constructed themes are presented in the following Results section to explain how cybersecurity appears in the Laurea case theses.

4. Results

Out of the fifteen theses selected to the sample, two were master's theses and thirteen bachelor's theses. The 13 bachelor's degree programme theses include seven from Safety, Security and Risk Management or its former curricula Security Management, and six from Business Information Technology. The two master's programme theses are from the Business Administration and Innovative Digital Services of the Future programmes. The included thesis projects were all commissioned, by private organizations, Laurea research, development, and innovation (RDI) projects, or state authorities.

Three of the sample theses are in some aspect related to the develop resilience and business continuity aiming. Four of the examined reports had technical foci and aimed to develop case organizations' activities and abilities to respond to future developments. The sample theses became sorted into seven categories according to their themes as shown in Figure 1.

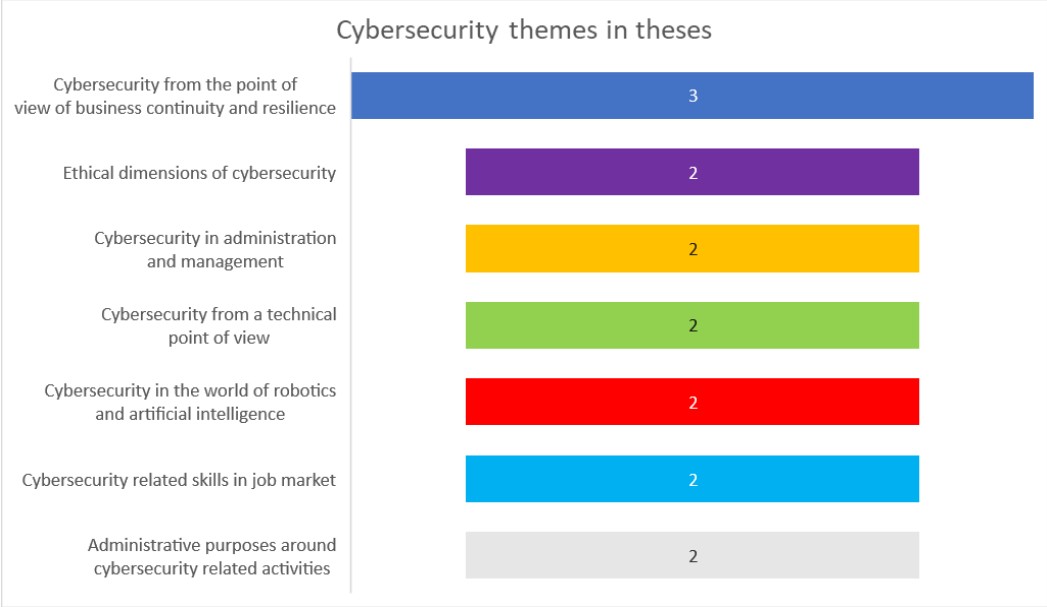


Figure 1. Seven themes identified from the theses.

These themes identified from these theses contain three theses in one category and two theses all the others. This Results section is structured according to these themes: Cybersecurity from the point of view of business continuity and resilience, Ethical dimensions of cybersecurity, Cybersecurity in administration and management, Cybersecurity from a technical point of view, Cybersecurity in the world of robotics and artificial intelligence, Cybersecurity related skills in job market, and administrative purposes around cybersecurity related activities.

4.1 Cybersecurity from the point of view of business continuity and resilience

The biggest category, with three theses, is on the theme of Cybersecurity from the point of view of business continuity and resilience, an IT security audit, a cybersecurity plan for the case company, and the protection of information assets.

(Reinikka, 2021) conducts an IT security audit on the office systems of a paper mill; a case study about an audit performed together with their partner organization following the framework provided by the Center of Internet Security (CIS) called CIS RAM (Risk Assessment Method). This framework is used to evaluate possibilities and impacts of vulnerabilities and threats identified in the risk assessment of the mill office systems.

(Gamoulos, 2021) executes a development project which creates a cybersecurity plan for the case company, a multimedia company that offers expertise and services that focus on website development, digital marketing, and video and film production. Interviews, a workshop, and risk assessment tools are used to collect data, and to put together a cybersecurity plan introducing e.g., the NIST framework, which is a voluntary framework of standards, guidelines, and best practices to manage cybersecurity risk.

(Koivuniemi, 2022) views in a research-based development project what cybersecurity is needed to protect the information assets of the commissioning organization. The main goal in the thesis is to increase resilience in the commissioning organization.

4.2 Ethical dimensions of cybersecurity

Two theses are on the theme exploring the ethical dimensions of cybersecurity. One thesis examines the contradictory values between biomedical and informational technology ethics, and the other offers guidelines for cybersecurity from the perspective of biomedical ethics.

(Kaukonen, 2021) as part of research, development, and innovation (RDI) activities of Laurea, generates enrichened information for a project examining contradictory in values between biomedical ethics and informational technology, according to the principles of the project, by interviews of a sample limited to elderly people.

(Hämäläinen, 2021) provides guidelines from a cybersecurity and biomedical ethics perspective to his commissioning project SHAPES, as part of Laurea RDI activities, utilizing Hevner's Design Science Research model in his work and results with respond to ethical questions related to privacy, autonomy, consent, and beneficence where ethical decision-making is required.

4.3 Cybersecurity in administration and management

There are two theses that investigate cybersecurity in administration and management. One thesis looks at administrative level process description of cyber situational awareness (CSA), and the other how to design and implement convergence for cybersecurity and physical security in railway operations.

(Hedeman, 2022) explores cybersecurity situational awareness. The results of the thesis show a difference between situational awareness and situational understanding. The main goal is to put together an administrative level process description of cyber situational awareness (CSA) for the commissioner of the thesis.

(Pacelli Queiroz Felix and Yusuf, 2021) conclude that designing and implementing cybersecurity and physical security convergence in railway operations. Core areas such as people, process, and technology, must be seamlessly integrated, suggesting that converging the physical and cybersecurity departments can improve the risk management ability and enhance the security and safety of the railway operations of their commissioning organization and its members.

4.4 Cybersecurity from a technical point of view

Two theses look at cybersecurity from a technical point of view. One thesis examines Internet of Things (IoT), and the other investigates information security threats e.g., cyberattacks and cyberthreats.

(Laiso, 2021) examines the Internet of Things (IoT) by conducting a content analysis literature related to the cyber security of IoT devices, open interviews with manufacturers, and a review of the cyber security materials produced by these manufacturers, highlighting the differences in levels of cyber security between different device types and manufacturers, and also showing connections between low price and low level of cybersecurity, and company size and capability of cybersecurity documentation.

The thesis by (Syrjälä, 2021) approaches the cyberworld through information security using search words that address information security threats, such as cyberattacks and cyberthreats. The main goal of the thesis is to produce useful information about information security threats of IP cameras, and so its aims are technical in nature.

4.5 Cybersecurity in the world of robotics and artificial intelligence

Two theses fall under the category of cybersecurity in the world of robotics and artificial intelligence. These two theses investigate how artificial intelligence and robotics can assist leadership the financial sector, and the cybersecurity of care robots in the fields of physical and mental care. New future risks are to be expected.

(Anttila and Rantanen, 2021) concentrate on how artificial intelligence and robotics can assist leadership. The thesis views cybersecurity from the point of view of what skills are needed in future management and leadership in the financial sector by performing qualitative research with data collection through interviews and offering a model profile for leadership in the financial sector, with an overview of the effects of digitalization on Finland's financial sector.

(Järvinen, 2021) examines cybersecurity of care robots in the fields of physical and mental care. The study shows how cybersecurity threats are associated with the same risks and threats as the use of other IT devices and robots. The results, however, also address the underlying human factors behind cybercrime, and showing that potential threats are remote access of care robots, espionage, and eavesdropping through network connectivity; new risks by artificial intelligence and machine learning are expected in the future.

4.6 Cybersecurity related skills in job market

Two theses address cybersecurity related skills in the job market. One of these studies shows a possible gap between what skills are provided by Finnish higher education providers and what Finland's work life is looking for, while the other analyses what are the most desirable skills for cybersecurity on the Irish job market.

(Luoma, 2021) uses qualitative research to analyse how cybersecurity skills gained from degree programmes relate on the Finnish labour market. The thesis analyses job advertisements that are published in different forums and compares the skills found in them with different the learning objectives of Finnish higher education degree programmes. The results show that there is a possible gap between what the education provides and what the working life is looking for.

(Walsh, 2021) analyses what are the most desirable skills for cybersecurity on the Irish job market. Relevant cybersecurity skills are divided into four categories: technical, situational awareness, problem solving, and sector specific skills. Results indicate that technical skills are the most descriptive, while sector specific skills are the most infrequently asked. This case study enrichens to the larger body of knowledge of the research project ECHO funded by the European commission.

4.7 Administrative purposes around cybersecurity related activities

Two studies examine cybersecurity from the perspective of administrative purposes around cybersecurity related activities. Both works enrichen the body of knowledge of the project ECHO as they look at Share Point portal models that could be used for data and information governance of the future ECHO network.

The study by (Lötjönen, 2021) promotes two possible models for a Share Point portal that could be used for data and information governance purposes for the ECHO network future organization. The aim to identify how to organize information governance in the cybersecurity network enrichens the body of knowledge of the project ECHO project.

The thesis study by (Koskula, 2021) covers project administration related information systems and examines possible uses of SharePoint in that connection. The connection to cybersecurity is that it too enrichens the body of knowledge of the project ECHO.

5. Conclusions

Based on the results of this case study cybersecurity is a topic that is being examined and/or developed in the theses written at Laurea University of Applied Sciences. The theme is discussed from several points of view and in multidisciplinary ways. The size of theme-based categories that can be recognized remains small during the sample period of this study. The theme-based categories show that the points of view that these thesis writers have taken towards cybersecurity vary significantly. The themes and their key features are presented in Table 1. This review mostly reflects interests of the students, but also the requirements of work life from industry, civil service, or Laurea RDI activities. This study helps understand what cybersecurity topics need to be addressed on in the classroom and to guide the selection of future thesis topics. This is a contribution to the scientific community and can be applied in other HEIs. The contribution to theory is a deeper understanding of how cybersecurity is seen in higher education.

Cybersecurity literature indicates that cybersecurity is everyone's concern and a topical issue, so it was expected to show in higher numbers of cybersecurity related thesis topics. However, after the search and looking into the theses found, it was somewhat surprising that not more theses would mention cybersecurity as a key concept or main field of interest.

Table 1. Cybersecurity themes and their key features

Cybersecurity theme	Key features
Cybersecurity from the point of view of business continuity and resilience	Risk management Continuity planning Asset protection
Cybersecurity in administration and management	Situational awareness Cybersecurity of processes
Ethical dimensions of cybersecurity	Ethically sustainable applications and cybersecurity solutions
Cybersecurity from a technical point of view	Internet of Things Network security Hardware and software security
Cybersecurity in the world of robotics and artificial intelligence	Artificial intelligence Robotics
Cybersecurity related skills in job market	Labour market requirements
Administrative purposes around cybersecurity related activities	Cybersecurity as a business or field of interest

It could be interesting to explore the theses that were excluded in this study to better understand why they appeared in the search. Reasons may vary and some of them may have appeared because they state that cybersecurity is excluded from the scope of their investigation. Many reports seem to have handled information security, and it would be especially interesting to see how they have defined the concept information security and what from their aspect is the role or definition of cybersecurity, and why it did not appear in their abstracts. This raises the question how well the concept cybersecurity is understood among higher education students.

Based on the findings, cybersecurity can be seen as a safe barrier around an organization's most valuable assets. When well organized and widely applied, cybersecurity measures may enhance business continuity. Figure 2 shows how cybersecurity could be seen as a circle around business continuity, which, according to literature, is a main target for every organization.

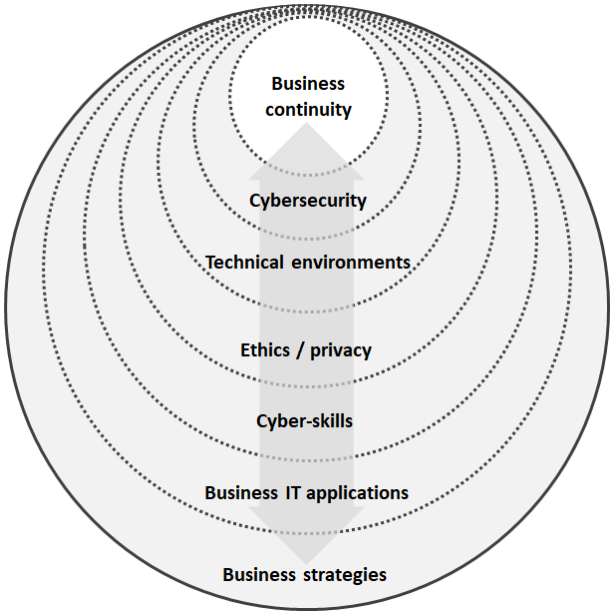


Figure 2. Model of the dimensions of cybersecurity.

As seen in Figure 2, the themes that were identified from the sample of this study recognize some dimensions of cybersecurity. Mainly, cybersecurity seems to be an essential support for business continuity. Secondly, comprehensive cybersecurity builds on appropriate cyber-skill levels and well working reliable technical environments. Neither of these alone provide security, they are needed together, and technical environments should be configured and used in ethical ways to ensure privacy when using business-IT applications. Cybersecurity becomes visible through business IT applications, which are cyber-physical in nature as they contain both digital and physical elements.

The model in Figure 2 also indicates that cybersecurity measures support business strategies by promoting business continuity, when applied throughout the entire organization and its processes. As some measures can increase the protection of some assets, they may at the same time cause harm to other assets or values, which need to be protected and the measures need to be in the right relation to the risks that they are planned to protect from.

It becomes essential to first expose any planned applications to the critical ethical discussion, before applying cybersecurity measures or solutions into the activities and processes of the organization. These ethical discussions should continue throughout organizational processes ongoingly also after the implementation of the applications and their cybersecurity measures.

Acknowledgements

This work was supported by the DYNAMO project, which has received funding from European Union's Horizon Europe research and innovation funding programme under the grant agreement no. 101069601.

References

Aaltola, K. and Taitto, P. (2019) 'Utilising Experiential and Organizational Learning Theories to Improve Human Performance in Cyber Training', *Information & Security: An International Journal*, 43(2), pp. 123–133. doi: 10.11610/isij.4311.

Anttila, A. and Rantanen, P. (2021) *Robotti digityön pomona : tekoäly ja robotiikka johtamisen apuna finanssialan työn murroksessa*. Available at: http://www.theseus.fi/handle/10024/502387 (Accessed: 17 February 2023).

Bowden, M. (2010) 'The Enemy Within', *The Atlantic*, 305(5), p. 14.

Burgess, J. P. (2010) *Handbook of New Security Studies*. Routledge.

Cavelty, M. D. (2010) 'Cyber-Security', in *The Routledge Handbook of New Security Studies*. 1st edn. London: Routledge, p. 328.

Conkova, M. (2013) 'Analysis of Perceptions of Conventional and E-Learning Education in Corporate Training', *Journal of Competitiveness*, 5(4), pp. 73–97. doi: 10.7441/joc.2013.04.05.

Craigen, D., Diakun-Thibault, N. and Purse, R. (2014) 'Defining Cybersecurity', *Technology Innovation Management Review*, 4(10), pp. 13–21.

Cupiał, M. *et al.* (2018) 'Information technology tools in corporate knowledge management', *Ekonomia i Prawo. Economics and Law*, 17(1), pp. 5–15.

Cybersecurity and Infrastructure Security Agency (CISA) (2019) *What is Cybersecurity?* Available at: https://www.cisa.gov/uscert/ncas/tips/ST04-001 (Accessed: 25 November 2022).

Gamoulos, C. (2021) *Developing a cybersecurity plan for the websites of case company X*. Bachelor's Thesis. Laurea UAS. Available at: http://www.theseus.fi/handle/10024/512312 (Accessed: 17 February 2023).

Hämäläinen, H. (2021) *Ethics of Cybersecurity and Biomedical Ethics – Providing Ethical Guidelines for the SHAPES Project*. Master's thesis. Laurea UAS. Available at: http://www.theseus.fi/handle/10024/500513 (Accessed: 17 February 2023).

Hedeman, E. (2022) *Kybertilannekuva valtioneuvoston kansliassa*. Bachelor's thesis|. Laurea UAS. Available at: http://www.theseus.fi/handle/10024/722709 (Accessed: 17 February 2023).

Hirsjärvi, S. *et al.* (2000) *Tutki ja kirjoita*. 6. uud. laitos. Helsinki: Tammi.

Hortovanyi, L. and Ferincz, A. (2015) 'The impact of ICT on learning on-the-job', *The Learning Organization*, 22(1), pp. 2–13. doi: 10.1108/TLO-06-2014-0032.

Im, T., Porumbescu, G. and Lee, H. (2013) 'ICT as a Buffer to Change', *Public Performance & Management Review*, 36(3), pp. 436–455. doi: 10.2753/PMR1530-9576360303.

Isidro-Filho, A. *et al.* (2013) 'Workplace learning strategies and professional competencies in innovation contexts in Brazilian hospitals', *BAR - Brazilian Administration Review*, 10(2), pp. 121–134. doi: 10.1590/S1807-76922013000200002.

Järvinen, M. (2021) *Hoivarobottien kyberturvauhkien kartoitus hoivarobotiikan asiantuntijoiden näkökulmasta*. Bachelor's thesis. Laurea UAS. Available at: http://www.theseus.fi/handle/10024/510631 (Accessed: 17 February 2023).

Kaspersky (2022) *What is Cyber Security?*, *www.kaspersky.com*. Available at: https://www.kaspersky.com/resource-center/definitions/what-is-cyber-security (Accessed: 25 November 2022).

Kaukonen, H. (2021) *Kyberturvallisuuden eettiset ulottuvuudet SHAPES-hankkeessa*. Bachelor's thesis. Laurea UAS. Available at: http://www.theseus.fi/handle/10024/426748 (Accessed: 17 February 2023).

Koivuniemi, M. (2022) *Jatkuvuudenhallinnan toimintamallin kehittäminen*. Bachelor's thesis|. Laurea UAS. Available at: http://www.theseus.fi/handle/10024/743772 (Accessed: 17 February 2023).

Koskula, S. (2021) *ECHO hallinto- ja johtamisen tietojärjestelmän kehitys ja SharePointin käytön tutkinta*. Bachelor's thesis. Laurea UAS. Available at: http://www.theseus.fi/handle/10024/511346 (Accessed: 17 February 2023).

Laiso, M. (2021) *Toimintamalli turvateknisten IoT-laitteiden kyberturvallisuustason arviointiin laite- ja järjestelmäasentajille*. Master's thesis. Laurea UAS. Available at: http://www.theseus.fi/handle/10024/500442 (Accessed: 17 February 2023).

Laurea University of Applied Sciences (2022) 'Laurean strategiset mittarit', *Organisaation johtaminen ja kehittäminen*. Available at: https://laureauas.sharepoint.com/sites/staffFin_johtaminenjakehittaminen/SitePages/koulutuksen%20strategiset%20mittarit%20-%20osa%201.aspx (Accessed: 30 May 2022).

Lexico (2022) *Definition of security*, *www.dictionary.com*. Available at: https://www.dictionary.com/browse/security (Accessed: 30 May 2022).

Limnéll, J., Majewski, K. and Salminen, M. (2014) *Kyberturvallisuus*. Jyväskylä: Docendo.

Lötjönen, L. (2021) *SharePoint portaali tukemaan ECHO-verkoston strategisen suunnittelun prosessia*. Bachelor's Thesis. Laurea UAS. Available at: http://www.theseus.fi/handle/10024/503346 (Accessed: 17 February 2023).

Luoma, S. (2021) *Cybersecurity skill development through degree programme relative to labor market*. Bachelor's thesis. Laurea UAS. Available at: http://www.theseus.fi/handle/10024/512744 (Accessed: 17 February 2023).

Merriam, S. B. and Tisdell, E. J. (2015) *Qualitative Research: A Guide to Design and Implementation*. 4th edn. San Francisco: John Wiley & Sons.

Mihalic, T. and Buhalis, D. (2013) 'ICT as a new competitive advantage factor – case of small transitional hotel sector', *Economic and Business Review*, 15(1), pp. 33–56.

Mohammed, S. (2015) 'An Introduction to Digital Crimes', *International Journal in Foundations of Computer Science & Technology*, 5(3), pp. 13–24. doi: 10.5121/ijfcst.2015.5302.

National Cyber Security Centre (2022) *cybersecurity - Glossary | CSRC*, *Cybersecurity*. Available at: https://csrc.nist.gov/glossary/term/cybersecurity (Accessed: 25 November 2022).

Pacelli Queiroz Felix, W. and Yusuf, H. M. (2021) *Measures needed to integrate physical and cyber security in railways infrastructure : organization X case study*. Bachelor's Thesis. Laurea UAS. Available at: http://www.theseus.fi/handle/10024/510112 (Accessed: 17 February 2023).

Payne, B. K. (2020) 'Defining Cybercrime', in Holt, T. J. and Bossler, A. M. (eds) *The Palgrave Handbook of International Cybercrime and Cyberdeviance*. Cham: Springer International Publishing, pp. 3–25. doi: 10.1007/978-3-319-78440-3_1.

Reinikka, M. (2021) *An Information security audit for a Finnish paper mill*. Bachelor's Thesis. Laurea UAS. Available at: http://www.theseus.fi/handle/10024/496026 (Accessed: 17 February 2023).

Ruoslahti, H. and Tikanmäki, I. (2022) 'Cybersecurity in Skills Development and Leadership', in *Future-proof Business - System Leadership Competences*. *3UAS: Future-proof Business - System Leadership Competences*, Virtual conference.

Salleh, K. *et al.* (2012) 'Learning and knowledge transfer performance among public sector accountants: an empirical survey', *Knowledge Management Research & Practice*, 10(2), pp. 164–174.

Shoemaker, D. and Conklin, W. A. (2011) *Cybersecurity: The Essential Body of Knowledge*. Boston, MA: Cengage Learning.

Škerlavaj, M., Dimovski, V. and Desouza, K. C. (2010) 'Patterns and structures of intra-organizational learning networks within a knowledge-intensive organization', *Journal of Information Technology*, 25(2), pp. 189–204. doi: 10.1057/jit.2010.3.

Syrjälä, T. (2021) *IP-valvontakameroiden tietoturvallisuus*. Available at: http://www.theseus.fi/handle/10024/466707 (Accessed: 17 February 2023).

Walsh, B. (2021) *Cyber security skill requirements from the Irish job market*. Bachelor's Thesis. Laurea UAS. Available at: http://www.theseus.fi/handle/10024/512741 (Accessed: 8 June 2022).

Zhao, F. and Kemp, L. (2013) 'Exploring individual, social and organisational effects on Web 2.0-based workplace learning: a research agenda for a systematic approach', *Research in Learning Technology*, 21. doi: 10.3402/rlt.v21i0.19089.

On the Software Architectures for fog-based Secure IOT Deployments

Christos Tselikis[1], Leandros Maglaras[2], Christos Douligeris[1] and Sarandis Mitropoulos[3]
[1]Department of Informatics, University of Piraeus, Piraeus 185 34, Greece
[2]School of Computer Science, Edinburgh Napier University, Edinburgh, UK
[3]Ionian University, Regional Development Department, Lefkada, Greece

ctselikis@unipi.gr
l.maglaras@naier.ac.uk
cdoulig@unipi.gr
sarandis@unipi.gr

Abstract: In this paper, we examine architectural designs for the support of demanding ad hoc IoT applications, such as industrial and large-scale IoTs. First, we examine the traditional software stack of nodes involved in centralized sensory applications. Then, we propose a highly distributed ad hoc architecture with increased node cooperation. Finally, we propose a secure fog-based hybrid model that offers optimizations with respect to performance and security and which facilitates the development of intelligent localized end-user applications with very strict latency requirements. In the three models that we examine we highlight operations at the routing layer and at the clustering sub-layer.

Keywords: IoT, Fog Computing, Security, Trust, AI/ML.

1. Introduction

Optimum architectural designs are essential for the successful deployment of secure Internet of Things (IoT) networks that are involved in modern use cases, such as industrial Internet of Things (IIoT) and massive IoT. It is well-known that time-critical IoT applications impose very strict requirements for latency, reliability and availability. Specifically, according to the IMT-2020 requirements (ITU 2017), the end-to-end latency in industrial IoT (e.g., robotic control in smart factories) should be less than one-millisecond. In addition, the massive IoT deployments generate *big data* which is not trivial to handle with legacy technologies.

For such time-sensitive and large-scale IoT applications, we identified a number of intensive requirements that can only be satisfied with the adoption of novel, efficient and effective automated solutions:

- There is an increasing demand for achieving even better network performance in support of time-sensitive applications. To satisfy very strict performance requirements, the fog architecture (Open Fog Reference Architecture 2017) can be the solution to effectively reduce the network response times.
- There is strong demand for novel applications with local characteristics which can assist to handle contemporary social needs, such as environmental, health-related, and transport-related. To this end, node cooperation is vital in order to be able to perform in-time, informed and automated decision-making at the local vicinity of interest (e.g., a parking lot, a shopping mall or a historical city sector).
- There is an intensive need to efficiently handle the *big data* inside mobile content delivery networks. This *Big data* is generated especially by the use of multimedia and geospatial applications, for example by vehicle passengers. Also, very large-scale sensory data is generated in the massive Internet of Things. Computing and storage capacity to perform automated AI/ML processing of the *big data* is vital to tackle this requirement.
- There is a need for adaptive network management. ML-based network design, resource management and control with AI/ML have already been successfully implemented in various networked IoT deployments.
- There is a strong need for stronger security (for example, certificate-based mutual authentication between devices as well as between users and the network) and for privacy enhancing mechanisms such as layered use of pseudonyms. We propose the integration of cooperation mechanisms into the wireless communications system in order to increase the array of attacks that can be handled effectively.

To tackle these challenges, the existing communications models and software architectures need optimization if not complete adaptation. As we will see in the next sections, we have to utilize and effectively integrate the

new computing models (such as fog computing) in the deployments of Internet of Things which nowadays expand at very large scale.

This paper is organized as follows. After this first introductory section, which identifies the contemporary challenges for modern IoT applications., Section 2 presents the work related to the topic of security and privacy in novel fog-based architectures used in wireless ad hoc networks. In Section 3, we analyze three architectures and provide the details of the software stacks of the nodes involved in ad hoc IoT applications. The fourth section presents the proposed security solutions, while Section 5 draws the main conclusions and states the directions for future work.

2. Related Work

In this section, we examine and present related work on fog-based security solutions that have been recently proposed. These works fall into several categories, ranging from pure fog computing to combinations integrating blockchain techniques and advanced authentication methods.

In (Vaquero 2014), based on the readings aggregated from the sensors, the reputations of the sensors are calculated and then these reputations are classified differentially with the use of adaptive rather than fixed reputation levels. Data sets are trusted only from high reputation nodes. Fortino et al. (Fortino 2020) examined several architectural options in the integration effort of IoT with fog computing. Security and privacy issues of fog-based architectures have been extensively examined recently in several survey papers (Mukherjee 2020), (Mukherjee 2017), (Rani 2022).

In (Alzoubi 2022), the authors analyzed several blockchain-based fog computing solutions and how these solutions can achieve a distributed and trusted system, identity management, secure data, reputation, and payment systems. According to their findings such convergence would take several years and a great deal of work to be accomplished. The authors in (Katsikogiannis 2018) propose a policy-based management scheme utilizing the Service Oriented Architecture (SoA) for machine-to-machine communications.

The secure authentication of devices and users in a complex system that consists of end nodes, fog nodes, cloud providers, and several hardware and software platforms is a rather difficult problem to solve. The solution must be kept simple and computationally efficient while at the same time offering a high level of security. In a recent article that focuses on the security of healthcare architectures, the authors propose an efficient and authenticated key establishment scheme that resists common attacks and reduces computation overhead (Li 2022). Focusing only on the authentication of end users, another recent work proposes a combination of honeywords and authenticators that can offer protection against several attacks, while keeping the process simple and platform-agnostic (Papaspirou 2021).

Fog computing is mostly related to data that are collected from the end devices and sent to the fog nodes for computation, aggregation or analysis. In several domains, like the healthcare sector, these data may be sensitive and sometimes it is crucial to be able to correctly classify them before deciding to send them to a nearby fog node or to the cloud. The authors in (Sarrab 2022) proposed a framework for the classification of streamed data according to their criticality level. This framework can be used in order to reassure better protection of sensitive data and can be applied to other domains. Privacy preservation of data in fog computing is a critical matter that has been extensively researched recently (Sarwar 2022).

3. Architectural Designs

In this section we examine three different architectural designs, namely the centralized, the fully distributed and the hybrid ones under IoT scenarios. We examine if and how each one of the three different software architectures can address the challenges we identified in our introduction regarding time-sensitive, large-scale and secure localized applications. In each case, we focus on the operations at the routing layer and at the clustering sub-layer.

3.1 Centralized architecture

We examine the centralized IoT model based on the use case of an environmental application. In this scenario, there exist sensor nodes that forward their readings (e.g., moisture reports) to a sink point which is the networked base station with Internet connectivity. The base station performs the main information processing requested by the sensory application, makes the results available to the Internet and also sends commands or queries to manage the deployed sensors. We assume here that the gateway runs a clustering application to hierarchically organize the sensor nodes in clusters.

Initially, the base station assigns to the registered wireless small devices (sensors, PDA/smart phones) their configuration settings over the radio interface (e.g., IEEE 802.15.4). Figure 1 shows the sensors' software stack consisting of the physical/node discovery, the routing and the application layers and also highlights the clustering module installed at the gateway node. Physical parameters include the environmental data, the sensor location, the velocity, the connectivity degree and the energy which along with each sensor's neighbour list are periodically forwarded and passed to the clustering module at the gateway node. The base station executes a number of clustering runs to determine the cluster head set which defines the sensor network structure. Once the network structure has been determined, the base station broadcasts the cluster head list to the sensors via the radio control channel. The sensors can now forward their readings encrypted to the known local cluster heads. The latter can use default routing to reach the base station in one hop. Distance vectors are usually preferred for ad hoc routing in multi-hop scenarios (Sheng 2013). Alternatively, geographical routing can be utilized (Aznaoui 2021).

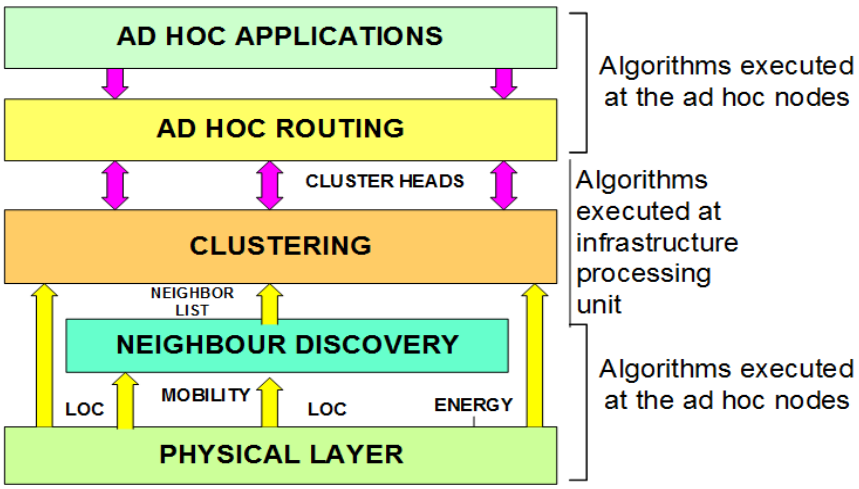


Figure 1: Software architecture for the centralised IoT model.

The above centralized design is more suitable for deploying static sensory applications of a small scale, such as, for example, monitoring of a small agricultural field. It is evident that this kind of processing at the local sink node cannot handle the *big data* that contemporary massive IoT applications generate nor can it address the dynamicity of mobile applications such as vehicle-to-vehicle image/video/data sharing scenarios.

3.2 Distributed cooperative architecture

The second architecture, shown in Figure 2, is a generic cross-layer framework that we propose for ad hoc IoT networks. Not all the components shown in Figure 2 are to be implemented at the network nodes; the actual implementations depend on the specific application's requirements, the capacity of the nodes and the network deployment choices available to the operator. This architecture differs from the centralized design in that now highly distributed interactions exist amongst peer ad hoc devices. The distributed components reside at the middleware layer providing a number of services to both end-user and management applications, including trust-security services, brokered publish-subscribe messaging services, service discovery, and energy and load management.

The devices now take autonomous decisions by participating in cooperative distributed protocols, such as in clustering protocols for the selection of local cluster heads or for the selection of local Certificate Authorities that establish trust. Every ad hoc node, due to the middleware, can become, for example, a local cluster head without the need of a gateway decision, or it may become a responsive intruder detector that broadcasts alerts in the event of malicious actions. Note that the malicious nodes, either external or internal, may launch various types of network attacks. Such attacks may target to harm a) the operations of the network protocols, b) the integrity of the data exchanged inside the network, and c) the users' privacy, mainly their identity, location and content of personal communications.

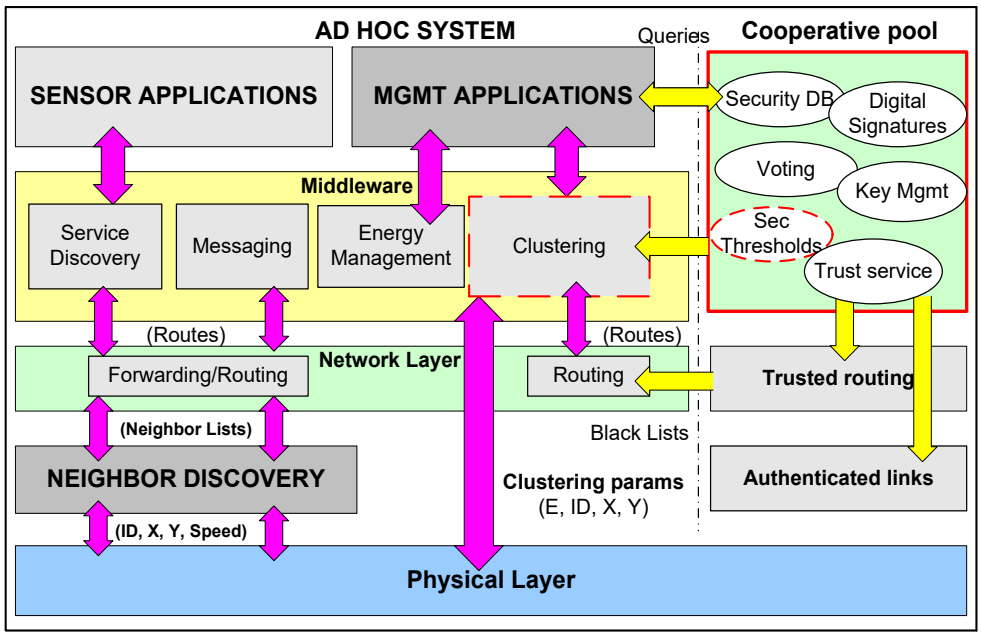


Figure 2: Software architecture for the distributed IoT model.

We enhance the IoT middleware layer with a pool of distributed components which are implemented as software libraries and which provide additional cooperative services to the end-devices. For example, a re-clustering process running at the ad hoc IoT plane can make a remote procedure call to the service of a distributed voting protocol which is based on security thresholds (Tselikis 2012). According to this scheme, the local nodes can securely reach an agreement regarding the identities of the IoT nodes that have the capacity to act as local Cluster Heads (or as local Certificate Authorities).

As another example of using the pool of cooperative components shown in Figure 2, the re-clustering procedure can trigger the execution of a Group Key Agreement protocol to establish, on-the-fly, a symmetric secret key. This key can be utilized as input to a symmetric algorithm to encrypt the messages exchanged inside an ad hoc group of nodes, such as industrial sensors or vehicles (Tselikis 2020). Alternatively, the routing protocol can call the services of a trust-based mechanism in order to resolve the next hop node based on the level of the nodes' reputation (Zahariadis 2010).

In our opinion, the distributed architecture shown in Figure 2 although is suitable for dynamic ad hoc IoT deployments, it lacks provisions for handling with analytics the big data generated in the massive IoT and also imposes computing overhead on the restricted IoT nodes. In addition, tight synchronization guarantees must be satisfied to successfully conduct the compound distributed operations present in this architecture.

3.3 Hybrid architecture

We propose the use of a hybrid architecture to meet the IoT requirements for small latency, big data handling, robust automated analytics processing with AI/ML, redundancy, scalability, privacy and security. It was made evident before that neither the centralized nor the distributed architecture can tackle the requirements in massive IoT and those in industrial environments. To this end, we propose to integrate the infrastructure-less wireless IoT networks with the cellular network adopting novel models, namely the fog computing paradigm and the disaggregated software-based architectural design of the Open Radio Access Network initiative (O-RAN Alliance 2023).

The fog architecture inserts a robust distributed layer, namely the fog tier, between the IoT devices and the centralized infrastructure at the cloud tier. The communication with fog nodes decreases the end-to-end latency and saves bandwidth since there is no need to contact the remote cloud for services. The end-devices generate a large number of data reports relating to local area information (e.g., images/video of road conditions). This data is hierarchically stored in the fog nodes or it can be optionally sent to the cloud.

In this context, new end-user applications can be designed and served by local servers installed at the vicinity of events of interest. For example, fog nodes installed at Road Side Units (RSU) may serve user queries regarding a free parking space (Ni 2017). On the other hand, local organizations or commercial entities can apply advanced processing to the data stored in the fog nodes in order to predict conditions and, therefore, make better

decisions (e.g., a transport entity can deduce knowledge for excessive pollution levels of a city area and then can re-route the traffic accordingly).

The IoT devices can communicate directly with each other or via the cellular network. In such a hybrid model, we can place the local servers which are envisaged in the fog model at the local infrastructure of the cellular network, i.e., at the 4G/5G base stations. Having more computational, communication and storage resources, the base stations can operate not only as local access points (traffic aggregators) but as intelligent estimators/predictors by utilizing AI/ML algorithms. The latter can run (or move) as virtualized/containerized components in the fog. In addition, in this way, processing-intensive tasks, such as intruder detection and routing optimizers, can be offloaded to the robust fog nodes.

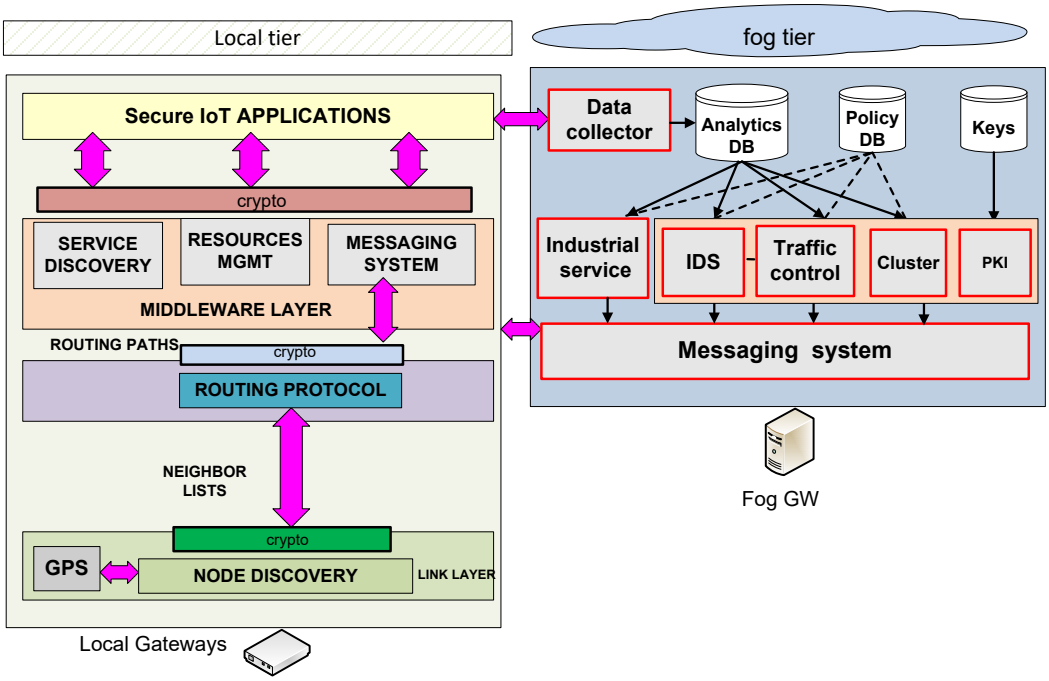


Figure 3: Fog-based hybrid IoT architecture.

As shown in Figure 3, with our proposal the network operator has the ability to utilize location-aware AI/ML-based controllers and analytics databases, as those are hierarchically envisaged in the Open RAN architecture, however at the fog tier. In more detail, an ML-based IDS component is shown in Figure 3. IDS agents monitor in real-time the end-user traffic collected and stored in a fog database and detect anomalies (e.g., DoS/DDoS attacks) or other suspicious user actions based on the classification made by a pre-trained ML model. The IDS agents also react in the event of such attacks by utilizing the messaging subsystem to broadcast alerts inside the network.

Further functionality can be achieved in case the agent interfaces with the Traffic Controller (TC). The traffic controller is another component used for the purposes of policy-based network management. This controller can act as a policy engine that loads the policies from a fog database and interfaces with the ad hoc messaging system. It is vital that TC can send (re)configuration commands to the IoT devices, or to other components that belong to the fog tier, based on the input fed by the IDS agent, as previously explained. If such an interface between the IDS agent and the TC is supported, the IDS agent, in the event of an attack, can trigger a signal to the TC containing information on the type of the launched attack. Then, the TC engine based on this input, can select new actions policies (security-wise or performance-wise) sending appropriate configuration commands to the devices and network elements.

The third architectural scheme can cope with a mixture of network models, e.g. with 4G/5G cellular IoT and with Vehicular to Infrastructure (V2I) applications. More specifically, it is proposed that the fog components and their interfaces shown in Figure 3 can be implementations compliant with the specifications of the Open RAN disaggregated architecture which introduces open and standardized interfaces between the network elements in the Radio Access Network.

4. Security

Fog nodes need to handle a large number of IoT devices or moving vehicles. In several cases, the users' data are exposed due to untrusted components residing on fog nodes, for example compromised nodes with malicious software. Privacy preservation is more challenging in fog computing since fog nodes that are in vicinity with the end users can collect sensitive data concerning the identity, communication pattern or coordinates. This information can be used for profiling and inference. Moreover, since fog nodes can be geographically distributed, centralized control becomes difficult.

In order to be able to secure our hybrid model we propose the combination of several solutions that include cryptographic methods, consistency controls, anti-spoofing mechanisms and intrusion detection systems. To achieve multi-fence protection, we adopted complementary protective solutions as suggested in (Mukherjee 2017). As shown in Figure 3, each ad hoc layer in the stack is protected with cryptographic primitives. Moreover, we exploit the fact that when location knowledge is available to the routing protocol, another layer of protection is offered against malicious nodes who broadcast faulty topology claims, given that the location and the identity of the nodes are cryptographically protected through digital signatures. In order to achieve this goal, we propose to have a fog-based PKI system that creates and stores the long-term keys for every IoT node. These keys are used to encrypt the sent messages and also to digitally sign them. Each node is certified by a Certificate Authority (CA) that issues a digital certificate that includes information about the location of the node.

The proposed model includes an intrusion detection system that constantly monitors, collects and aggregates the packets that are exchanged among nodes in its proximity. When the behaviour of the local neighbourhood of the IDS differs from the normal one that the IDS 'knows' the IDS sends alarms to the administrator of the system that can further investigate the situation similar to the one presented in (Ferrag 2019).

IoT devices must be able to verify whether the fog nodes are secure and, therefore, trust should be established. Our model incorporates the idea of clustering and voting. The nodes create clusters based on their proximity using distributed algorithms like the ones presented in (Maglaras 2012) and elect as cluster head the one that is more stable among the candidates. The cluster head can take over several actions in the neighbourhood, can act as a data collector, or even perform intrusion detection processes among others.

5. Conclusions

In this paper, we described traditional and new proposed software architectures for IoT deployments. We presented operational examples involving some of the ad hoc communications protocols (clustering and routing) that are involved in time-sensitive IoT application scenarios.

The centralized architecture, which is still being utilized in many sensory applications, is suitable for small-scale and mainly static wireless sensor networks. The distributed peer-to-peer architecture has attracted strong interest in the previous years and many middleware systems have been developed and utilized so far in commercial IoT deployments. However, in our opinion, it is doubtful not only whether the peer-to-peer model can tackle the performance and security requirements of the URLLC Use Case, but whether it can satisfy the need to efficiently handle the *big data* that modern large-scale multimedia applications create. In order to strengthen the security guarantees, we proposed cooperative enhancements in the case of the distributed peer-to-peer architecture.

Finally, we proposed a hybrid software framework that integrates the infrastructure-less ad hoc networks with the fog computing paradigm allowing the design of novel localized applications with stronger security (resilience to attacks) and better performance (smaller communications latency). We integrated in this fog-based model certain AI/ML components, namely the IDS agent and the Traffic Controller. By exploiting the robust computing and storage resources available at the fog tier, the IDS can run ML models which facilitate the detection of traffic anomalies inside IoT networks. In addition, the IDS component can be responsive in the event of attacks by applying a new configuration of the things tier.

In future work, we shall investigate further the integration path between wireless IoT networks and 5G cellular networks applying architectural concepts used in the Open RAN specifications. We could further investigate the integration of blockchain into the proposed model in order to further enhance privacy preservation and the integrity of the information but having in mind also the need for keeping a high quality of service in terms of latency and throughput.

Acknowledgements

This work has been partially supported by the European Union through project CybereSecPro (Project No. 101083594).

References

Alzoubi, Y. I., Al-Ahmad, A. and Kahtan, H. (2022) “Blockchain technology as a Fog computing security and privacy solution: An overview”, Computer Communications, Volume 182, pp. 129-152.

Aznaoui H., Ullah A., Raghay S., Aziz L. (2021) “Advanced GAF routing protocol using the goal attainment method in WSN”, NISS (ACM): 11:1-11:7.

Ferrag M. A. and Maglaras L. (2019) “Deliverycoin: An ids and blockchain-based delivery framework for drone-delivered services,” Computers, vol. 8, no. 3, p. 58.

Fortino G. et al. (2020) “Trust and Reputation in the Internet of Things: State-of-the-Art and Research Challenges”, IEEE Access, Volume 8, pp. 60117-60125.

ITU Report (2017), ITU-R M.2410. Available at: https://www.itu.int/pub/R-REP-M.2410, [accessed 28-03-2023].

Katsikogiannis G. et al. (2018) “A policy-aware Service Oriented Architecture for secure machine-to-machine communications”, Journal of Ad Hoc Networks, 80, 208, pp. 70-80.

Li, X. et al. (2022) “An efficient and authenticated key establishment scheme based on fog computing for healthcare system”, Frontiers of Computer Science, 16(4), pp. 1-12.

Maglaras L. A. and Katsaros D. (2012) “New measures for characterizing the significance of nodes in wireless ad hoc networks via localized path-based neighbourhood analysis”, Social Network Analysis and Mining, Volume 2, no. 2, pp. 97–106.

Mukherjee M. et al. (2020) “Security and privacy issues and solutions for fog,” Fog and Fogonomics: Challenges and Practices of Fog Computing, Communication, Networking, Strategy, and Economics, pp. 353– 374.

Mukherjee M. et al. (2017) “Security and Privacy in Fog Computing: Challenges”, IEEE Access, Volume 5, pp. 19293-19304.

Ni J. et al. (2017) “Security, Privacy and Fairness in Fog-based Vehicular Crowdsensing”, IEEE Communications Magazine, pp. 146-152.

Open Fog Reference Architecture for Fog Computing, produced by the OpenFog Consortium Architecture Working Group. Available at:
https://site.ieee.org/denver-com/files/2017/06/OpenFog_Reference_Architecture_2_09_17-FINAL-1.pdf
[accessed 28-03-2023].

O-RAN Alliance. Available at: https://www.o-ran.org/ [accessed 28-03-2023].

Papaspirou, V. et al. (2022) “Security Revisited: Honeytokens meet Google Authenticator”, *7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, Ioannina, Greece, pp. 1-8.

Rani, S., Kataria, A. and Chauhan, M. (2022) “Fog computing in industry 4.0: Applications and challenges—A research roadmap” Energy Conservation Solutions for Fog-Edge Computing Paradigms, Springer, Singapore pp. 173-190.

Sarrab, M. and Alshohoumi, F. (2022) “Assisted Fog Computing Approach for Data Privacy Preservation in IoT-Based Healthcare”, Security and Privacy Preserving for IoT and 5G Networks, Springer, Cham, pp. 191-201.

Sarwar, K. et al. (2022) “Efficient privacy-preserving data replication in fog-enabled IoT”, Future Generation Computer Systems, Volume 128, pp. 538-551.

Sheng, Z. et al. (2013) "A survey on the ietf protocol suite for the internet of things: standards, challenges, and opportunities", IEEE Wireless Communications, vol. 20, no. 6, pp. 91-98.

Tselikis, C. et al. (2012) "Degree-Based Clustering Algorithms for Wireless Ad Hoc Networks Under Attack," IEEE Communications Letters, Volume 16, no. 5, pp. 619-621.

Tselikis C. et al. (2020) “On the conference key distribution system with user anonymity”, Journal of Information Security and Applications, Volume 54: 102556.

Vaquero L. M. and Rodero-Merino L. (2014) “Finding Your Way in the Fog: Towards a Comprehensive Definition of Fog Computing,” ACM SIGCOMM Computer Communications Rev., vol. 44, no. 5, pp. 27–32.

Zahariadis T. et al. (2010) "Trust management in wireless sensor networks", European Transactions on Telecommunications, 21.4: pp. 386-395.

Towards Norms for State Responsibilities Regarding Online Disinformation and Influence Operations

Brett van Niekerk[1] and Trishana Ramluckan[2,3]
[1] Durban University of Technology, South Africa
[2] University of KwaZulu-Natal, South Africa
[3] Educor Holdings, South Africa
brettv@dut.ac.za
ramluckant@ukzn.ac.za

Abstract: The Internet has provided a global mass communication system, and in particular social media technologies began a social revolution for the public sphere. However, these platforms have been exploited for the purposes of influence operations and disinformation campaigns to hinder or subvert national decision-making processes by affecting the policy makers, voters, or swaying general public opinion. Often this is achieved through manipulative means falling within a grey area of international and constitutional systems. Existing proposed normative frameworks for responsible state behaviour in Cyberspace have tended to focus on cyber operations. While online influence operations are recognised as a concern, they were not explicitly discussed in the frameworks, resulting in knowledge gaps related to countering influence operations and disinformation. There is a growing narrative that influence operations and disinformation campaigns are a cyber security issue and nations sometimes include legislation related to disinformation in cyber security. This indicates that existing cyber norms can be used to guide the development of norms for addressing disinformation and influence operations. This paper aims to propose a normative framework for state responsibility relating to influence operations emerging from thematic analysis of existing cyber norms and research on mitigating influence operations.

Keywords: Cyber diplomacy, Cyber norms, Disinformation, Influence operations, Information warfare

1. Introduction

Disinformation and coordinated influence operations, often with a strong online component, have been receiving attention since the alleged attempts to influence the 2016 US presidential election. During the COVID-19 pandemic there was an increase in disinformation. While disinformation had been recognised as a concern, limited coordinated international recognition of the problem with a view of mitigating disinformation have been established. Whilst the United Nations General Assembly (UNGA, 2021) passed a resolution in 2021, this does not have the maturity of 'traditional' cybersecurity, which has had multiple Groups of Governmental Experts (GGEs) with a second Open Ended Working Group (OEWG) at the time of writing and a proposed programme of action (Digital Watch, 2023).

Disinformation can be defined as the intentional spreading of false or misleading information, whereas misinformation is unintentional, and malinformation is the use of correct information used for misleading purposes (Khan, 2021). Influence operations are then considered the use of these techniques to manipulate and audience to support the goals of an actor, which may be a foreign actor seeking to subvert processes in another nation (Pamment, 2022).

This paper aims to begin the process of establishing norms for addressing disinformation and influence operations by analysing existing cyber norms and international documents on disinformation. Thematic analysis is used to identify key themes that the proposed normative framework should be based upon. The paper contributes an initial framework of norms for state responsibility in countering, conducting, and responding to online disinformation and influence campaigns, based upon the key themes identified in the document analysis.

Section 2 presents a background to facets of disinformation and influence operations, the relationship with cybersecurity, and cybersecurity norms. Frameworks and legal aspects for disinformation and influence operations at both international and national levels are discussed in Section 3. The proposed norms are presented in Section 4, followed by the conclusion in Section 5.

2. Literature Review

2.1 Major disinformation and influence operations

The major event that raised international concern regarding disinformation and foreign influence operations was the 2016 US presidential election, where allegations of interference resulted in a number of political investigations and processes. The actors used a variety of techniques, including using adverts on social media

platforms. Social data used to design manipulative messaging was linked to Cambridge Analytica and was also found to have influence the BREXIT referendum. Similarly, a private firm, Bell Pottinger, was at the centre of a scandal where disinformation was used to increase racial tensions in South Africa to distract from accusations of political wrongdoing against the president (Ramluckan, Wanless and van Niekerk, 2019; Ramluckan and van Niekerk, 2019).

Disinformation was prevalent in Ukraine during both the anti-government protests and the annexation of Crimea in 2014, and evolved during the 2022 Russo-Ukraine conflict to include the use of deepfakes (Roache *et al*., 2022). Disinformation again become an international concern due to its prevalence during the COVID-19 pandemic, with conspiracy theories and political blaming. In addition, a nation used 10 000 social media bots and compromised accounts for propaganda purposes (Allen-Ebrahimian, 2020).

Baines and Jones (2018) and Rid (2020) provide historical examples of election interference and influence operations, respectively. While they demonstrate that the concept is not new, the introduction of social media and cyber elements have evolved the practice dramatically. While many key examples focus on Europe, other areas are also affected by foreign influence operations. For example, approximately 60 percent of the over 50 recorded disinformation operations in Africa are considered to originate externally, with 16 attributed to Russia (Africa Center for Strategic Studies, 2022).

2.2 Influence operations versus cyber operations

Both influence operations and cyber operations both emerged out of the concept of information warfare, which encompassed military psychological operations and computer network operations; these were extended outside of a conflict situation to information operations. The terminology has evolved, where information operations and information warfare are often used synonymously with influence operations; for example, Stengel's (2019) book *Information Wars: How We Lost the Global Battle Against Disinformation and What We Can Do About It*, focuses purely on disinformation as the title suggests. Likewise, the term 'cyber operations' has become far more common and largely replaced the use of 'computer network operations'. Rid's (2020) book, *Active Measures: The Secret History of Disinformation and Political Warfare*, considers a cyber security incident as an influence operation: a group calling themselves the Shadow Brokers were selling cyber security tools stolen from the U.S. National Security Agency online; however, the narrative surrounding this appeared to be an influence operation to embarrass the agency as the tools were eventually released openly on the Internet.

Gleicher (20221; 2022b) indicates that there are similarities between cyber security and defending against disinformation and influence operations: both situations are adversarial, and the threats evolve to circumvent defences. Within the cyber security domain, cybercriminals offer specialist services for hire, and this has translated to services for hire for influence operations by relevant threat actors (Gleicher, 20221; 2022b). It is reported that a North African cybercriminal shifted the focus of operations from cybercrime to distributing online propaganda (Lunghi, 2017). Account compromises, a common occurrence in the cyber security domain, is used to gain access to platforms with which to distribute and amplify influence operation narratives (Gleicher, 20221; 2022b), illustrating a cross-over between cyber security.

The Disinformation Analysis & Response Measures (DISARM) Framework provides two key frameworks describing the disinformation creation tactics, techniques and procedures (TTPs) by threat actors and mitigation TTPs by defenders (DISARM Foundation, n.d.; Terp and Breuer, 2022). This is modelled on the MITRE ATT&CK and D3FEND frameworks for 'traditional' cyber security. In addition, a STIX (Structured Threat Information eXpression) capability is available for threat information sharing (DISARM Foundation, n.d.), as is common with cyber security. The European Union Agency for Cybersecurity and EEAS (ENISA and EEAS, 2022) explicitly maps components of the MTRE ATT&CK framework to the DISARM framework; however, it is indicated that the DISARM framework is aligned to describing campaigns, whereas the ATT&CK framework is better at describing a cyber security incident. ENISA and EEAS (2022) indicates important components of cyber security that are relevant to foreign information manipulation and interference: cyber operations can be used for developing content (deep fakes or hack and leak operations) to support a narrative, or can compromise accounts and infrastructure to provide a platform for influence operations and dissemination; from a defensive perspective, cyber security techniques can be used to investigate the technical aspects to influence operations and aid in attribution. These aspects echo that of Gleicher mentioned above.

One of the strategies proposed to mitigate influence operations and disinformation is a 'whole of society approach' (Gleicher, 20221; 2022b), similar to the multi-stakeholder approaches in cyber security and the UN OEWG. Wanless and Shapiro (2022) echo a collaborative approach, advocating for research into disinformation

and influence operations to be based on a model such as the European Centre for Nuclear Research (CERN); currently individual projects are funded, resulting in replication of infrastructure for each project, whereas a conglomerate of nations could support global researchers with a single large-scale project to maximise economy of scale.

Due to the volume of possible incidents, automation is required for detecting both cybersecurity and disinformation threats, with artificial intelligence (AI) and machine learning playing a major role. While AI and machine learning have been successful within the cybersecurity space (Crowley, 2022; Fein and Stocker, 2022), it has been reported that AI is limited in detecting disinformation and influence operations (U.S. Federal Trade Commission, 2022; Kelley, 2022). Currently AI is not able to adequately detect nuances of human speech, and the available datasets used to train algorithms are not yet diverse enough, and technological solutions alone is not enough to address the problem (Kelley, 2022; IEEE Computer Society, 2021). Therefore, it is important to have an international normative framework to guide a common approach to addressing disinformation and influence operations. The next section provides and overview of norms within the cybersecurity domain which is used in conjunction with the legal aspects covered in Section 3 as a basis to develop norms for mitigating disinformation in Section 4.

2.3 Norms for state responsibility in Cyberspace

The Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security in the 2013 report indicated that International Law applies to Cyberspace, and in a 2015 report proposed 11 norms for responsible state behaviour in Cyberspace (Brown, Esterhuysen, and Kumar, 2019; United Nations General Assembly, 2015):

a) "Consistent with the purposes of the United Nations, including to maintain international peace and security, States should cooperate in developing and applying measures to increase stability and security in the use of ICTs and to prevent ICT practices that are acknowledged to be harmful or that may pose threats to international peace and security" (Brown, Esterhuysen, and Kumar, 2019: 4).

b) "In case of ICT incidents, States should consider all relevant information, including the larger context of the event, the challenges of attribution in the ICT environment and the nature and extent of the consequences" (Brown, Esterhuysen, and Kumar, 2019: 5).

c) "States should not knowingly allow their territory to be used for internationally wrongful acts using ICTs" (Brown, Esterhuysen, and Kumar, 2019: 5).

d) "States should consider how best to cooperate to exchange information, assist each other, prosecute terrorist and criminal use of ICTs and implement other cooperative measures to address such threats. States may need to consider whether new measures need to be developed in this respect" (Brown, Esterhuysen, and Kumar, 2019: 5).

e) "States, in ensuring the secure use of ICTs, should respect Human Rights Council resolutions 20/8 and 26/13 on the promotion, protection and enjoyment of human rights on the Internet, as well as General Assembly resolutions 68/167 and 69/166 on the right to privacy in the digital age, to guarantee full respect for human rights, including the right to freedom of expression" (Brown, Esterhuysen, and Kumar, 2019: 5).

f) "A State should not conduct or knowingly support ICT activity contrary to its obligations under international law that intentionally damages critical infrastructure or otherwise impairs the use and operation of critical infrastructure to provide services to the public" (Brown, Esterhuysen, and Kumar, 2019: 6).

g) "States should take appropriate measures to protect their critical infrastructure from ICT threats, taking into account General Assembly resolution 58/199 on the creation of a global culture of cybersecurity and the protection of critical information infrastructures, and other relevant resolutions" (Brown, Esterhuysen, and Kumar, 2019: 6).

h) "States should respond to appropriate requests for assistance by another State whose critical infrastructure is subject to malicious ICT acts. States should also respond to appropriate requests to mitigate malicious ICT activity aimed at the critical infrastructure of another State emanating from their territory, taking into account due regard for sovereignty" (Brown, Esterhuysen, and Kumar, 2019: 6).

i) "States should take reasonable steps to ensure the integrity of the supply chain so that end users can have confidence in the security of ICT products. States should seek to prevent the proliferation of malicious ICT tools and techniques and the use of harmful hidden functions" (Brown, Esterhuysen, and Kumar, 2019: 6).

j) "States should encourage responsible reporting of ICT vulnerabilities and share associated information on available remedies to such vulnerabilities to limit and possibly eliminate potential threats to ICTs and ICT-dependent infrastructure" (Brown, Esterhuysen, and Kumar, 2019: 7).

k) "States should not conduct or knowingly support activity to harm the information systems of the authorized emergency response teams (sometimes known as computer emergency response teams or cybersecurity incident response teams) of another State. A State should not use authorized emergency response teams to engage in malicious international activity" (Brown, Esterhuysen, and Kumar, 2019: 7).

The wording of the above norms is generic enough that most of the norms are applicable, particularly as they consider malicious use of ICTs that could affect international security, and this is not necessarily limited to 'traditional' cyber-operations. The concepts of assistance, information exchange, protecting human rights, and not conducting malicious activity will all be applicable to disinformation as well as cyber-operations. Where the norms are particularly specific and technical, they may not be relevant to disinformation scenarios, such as protecting emergency response teams and core infrastructure. In addition to the GGEs an Open-Ended Working Group process was established with a key initiative being the inclusion of multi-stakeholder engagement, which has had challenges but has also resulted in beneficial engagement (Digital Watch, 2023).

The Global Commission on the Stability of Cyberspace (GCSC, 2019) proposed eight norms:

1. "State and non-state actors should neither conduct nor knowingly allow activity that intentionally and substantially damages the general availability or integrity of the public core of the Internet, and therefore the stability of cyberspace." (p. 21)
2. "State and non-state actors must not pursue, support or allow cyber operations intended to disrupt the technical infrastructure essential to elections, referenda or plebiscites." (p. 21)
3. "State and non-state actors should not tamper with products and services in development and production, nor allow them to be tampered with, if doing so may substantially impair the stability of cyberspace." (p. 21)
4. "State and non-state actors should not commandeer the general public's ICT resources for use as botnets or for similar purposes." (p. 21)
5. "States should create procedurally transparent frameworks to assess whether and when to disclose not publicly known vulnerabilities or flaws they are aware of in information systems and technologies. The default presumption should be in favour of disclosure." (p. 21)
6. "Developers and producers of products and services on which the stability of cyberspace depends should (1) prioritize security and stability, (2) take reasonable steps to ensure that their products or services are free from significant vulnerabilities, and (3) take measures to timely mitigate vulnerabilities that are later discovered and to be transparent about their process. All actors have a duty to share information on vulnerabilities in order to help prevent or mitigate malicious cyber activity." (pp. 21-22)
7. "States should enact appropriate measures, including laws and regulations, to ensure basic cyber hygiene." (p. 22)
8. "Non-state actors should not engage in offensive cyber operations and state actors should prevent such activities and respond if they occur." (p. 22)

These eight norms are more technical in nature, as the intention was to focus on the public core of the Internet, ensuring the correct technical functioning of the Internet. However, some of the norms contain concepts that can be adapted to the case of influence operations, such as election interference, using public ICT resources for botnets (which could be used for distributing disinformation), implementing appropriate measures for cyber hygiene, and restricting malicious behaviour by non-state actors.

The Paris Call (2018) comprises of 9 principles:

1. "Protect individuals and infrastructure: Prevent and recover from malicious cyber activities that threaten or cause significant, indiscriminate or systemic harm to individuals and critical infrastructure."
2. "Protect the Internet: Prevent activity that intentionally and substantially damages the general availability or integrity of the public core of the Internet."
3. "Defend electoral processes: Strengthen our capacity to prevent malign interference by foreign actors aimed at undermining electoral processes through malicious cyber activities."
4. "Defend intellectual property: Prevent ICT-enabled theft of intellectual property, including trade secrets or other confidential business information, with the intent of providing competitive advantages to companies or commercial sector."
5. "Non-proliferation: Develop ways to prevent the proliferation of malicious software and practices intended to cause harm."
6. "Lifecycle security: Strengthen the security of digital processes, products and services, throughout their lifecycle and supply chain."
7. "Cyber hygiene: Support efforts to strengthen an advanced cyber hygiene for all actors."
8. "No private hack back: Take steps to prevent non-State actors, including the private sector, from hacking-back, for their own purposes or those of other non-State actors."
9. "International norms: Promote the widespread acceptance and implementation of international norms of responsible behaviour as well as confidence-building measures in cyberspace."

Of the nine principles, some are directly relevant and some are partially relevant to influence operations. Concepts such as cyber hygiene and preventing election interference are again present. Non-proliferation can be applied to tools or capacity to conduct influence operations. Principle 9 should be of particular note: it explicitly supports normative processes, which therefore provides motivation for considering norms for mitigating influence operations.

3. Laws and Frameworks related to Disinformation and Influence Operations

3.1 International law and frameworks

In December 2021 the UNGA passed a resolution on *Countering disinformation for the promotion and protection of human rights and fundamental freedoms*, which recognises the harms disinformation can cause, and encourages civil society, media organisation (including social media) and states to take measures to mitigate the impacts of disinformation within the boundaries of human rights (UNGA, 2021). The resolution also recognises a special report to the UN Human Right Council on *Disinformation and freedom of opinion and expression* (Kahn, 2021). A key aspect that was noted in the report is that under human rights freedom of expression is only limited in that it should not infringe upon another's rights, therefore the spreading of disinformation is ultimately protected by human rights (Kahn, 2021); similarly, Schmitt (2017) indicates that "propaganda does not constitute a prohibited intervention as it is not coercive in nature", therefore, unless propaganda explicitly incites war or civil unrest in another country, it is unlikely to constitute a wrongful act.

Wanless and Shapiro (2021) point out that the co-sponsors of the above-mentioned UN resolution all have poor rankings in terms of corruption, media freedom and democracy; they suggest that authoritarian regimes may benefit from international norms that provide them a mechanism for controlling the information domain within their territories. Stengel (2019) indicates a similar dilemma: democracies allow for open debate; therefore a 'fringe' idea needs to have an opportunity to respond to critics, thereby giving the ideas a platform and a degree of credibility. By comparison, authoritarian regimes will simply sensor information. Consequently, democratic processes enable the spread on disinformation (Stengel, 2019).

Pamment (2022) in a publication by the NATO Strategic Communications Centre of Excellence considers both the capabilities and actions required to counter disinformation and influence operations. For the capability, four key considerations are proposed: objectives, indicators, process maturity and risk assessment; of note is the consideration of norms under the 'objectives' category. Mitigation measures include content correction (fact checking) and public resilience (media literacy), analysis and identification, strategic communication, intelligence, and security policies. In addition, the importance of partnerships and professional development are highlighted (Pamment, 2022). ENISA and EEAS (2022) as mentioned above provide a more technical focus, and

suggest partnerships between the cybersecurity and disinformation/influence operations communities, increasing the quality and quantity of disinformation incident information, and capacity building amongst the various groups. They proposed the DISARM framework (DISARM Foundation, n.d.; Terp and Breuer, 2022), which will give consistent reporting of key information on detected disinformation and influence campaigns to improve collaboration and monitoring. Gleicher (2022a; 2022b) proposes an 'Adversarial Design Cycle' to mitigate disinformation and influence campaigns, with four main stages: threat ideation to assess threats and trends to predict future threats and design countermeasures, threat disruption to monitor and mitigate threats, improved defences to increase resilience, and partnerships to facilitate a whole-of-society approach.

In the above-mentioned documents, key themes emerge: the need for partnerships and collaboration, and capacity building. The Carnegie Endowment for International Peace's (2023) Partnership on Countering Influence Operations (PCIO) is aimed at collaborative research and information sharing. The Atlantic Council's Digital Forensics Research Lab (DFRLab, 2023) runs the Digital Sherlocks programme, which is also research-focussed and has initiatives aimed at capacity building and collaboration.

3.2 National laws and processes

Gleicher (2022a; 2022b) suggests a whole-society approach and that governments should introduce regulations to help mitigate disinformation. In Africa some countries did this by introducing the criminalisation of COVID disinformation in 2020; however, there were concerns that national laws criminalising disinformation may allow totalitarian and authoritarian states to abuse the laws to censor information criticising the government or limit the spread of opinions and information that do not align to an official government position (Hodgson, Farise, and Mavedzenge, 2020; Moyo, 2020). For example, in 2017 a US journalist was arrested in Zimbabwe for allegedly insulting former president Robert Mugabe (Burke, 2017). In 2022 Russia outlawed the use of 'war' and 'invasion' in reference to what was termed a 'special military operation' in Ukraine (Simon, 2022).

The implementation of national legislation therefore exhibits a challenge in reducing disinformation while protecting the freedom of expression; as indicated in Section 3.1, the freedom of expression is protected even if the content expressed in false. Given the severity of disinformation and influence operations, there appears to be the need for alignment to international humanitarian law to strengthen the required responsibilities related to freedom of expression in order to limit the intentional spread of disinformation. In addition, nations put pressure on social media organisations to detect and limit the spread disinformation during the pandemic (de Wet, 2020; Gold, 2020). Other initiatives included fact checking, such as the Real411 platform by the Media Monitoring Africa (MMA, 2023), Viral Facts Africa (2023), and multiple agencies in Europe (López-Marcos and Vicente-Fernández, 2021). Kyriakidou, Cushion, Hughes, and Morani (2022) indicate that while there is a recognised need for fact checkers in the UK, there was low awareness amongst the participants of the existing fact checkers. López-Marcos and Vicente-Fernández (2021) indicate that fact checking in the UK was more like a business, whereas in Spain it was conducted by NGOs. This indicates multiple models that fact checking can be deployed in a nation, and also highlights the benefits of multi-stakeholder involvement in mitigating disinformation.

4. Proposed Normative Framework for State Responsibilities on Online Disinformation and Influence Operations

This section provides thematic analysis of the above norms, principles, and documents and indicates the relevance to disinformation and influence operations. These themes form the basis for proposing a normative framework for influence operations. Table 1 provides an indication of the relevance of existing cyber norms to disinformation.

Table 1: Relevance of Existing Norms and Principles to Disinformation and Influence Operations

Norm			Directly Relevant	Indirectly Relevant	Not Relevant
2015 GGE	a)	Cooperation for implementing measures		X	
	b)	Attribution		X	
	c)	Not allow wrongful acts in territory	X		
	d)	Information exchange		X	
	e)	Respect human rights	X		
	f)	Do not conduct or support ICT activity contrary to international obligations		X	
	g)	Protect critical infrastructure			X
	h)	Respond to requests for assistance		X	
	i)	Integrity of the supply chain			X
	j)	Reporting of vulnerabilities			X
	k)	Do not harm CERTs			X
GCSC	1	No activities that damage public core of the Internet			X
	2	No activities that damage the technical infrastructure for elections		X	
	3	Do tampering with products			X
	4	Do not commandeer ICT resources for botnets		X	
	5	Vulnerability disclosure		X	
	6	Developer obligations			X
	7	Implement measures and laws for cyber hygiene		X	
	8	No non-state offensive operations		X	
Paris Call	1	Protect individuals and infrastructure		X	
	2	Protect the Internet			X
	3	Defend electoral processes		X	
	4	Defend intellectual property			X
	5	Non-proliferation		X	
	6	Lifecycle security			X
	7	Cyber hygiene		X	
	8	No private hack back			X
	9	International norms		X	

Norm e) can apply directly to the case of disinformation, as the wording is broad enough that it does not necessarily include technical subversion of infrastructure, but the use of ICTs in general; norm c) can be incorporated into mitigating non-state activity within a territory. Due to the concerns around the implementation of legislation, this is excluded from the proposed norms. In addition, as disinformation and influence operations do not breach international law unless there is a specific impact (e.g. interference on national processes), a generic norm to refrain from such operations is not suitable, and can be covered by not interfering in national processes.

The proposed norms for responsible state behaviour regarding disinformation and influence operations are:

1. States "should respect Human Rights Council resolutions 20/8 and 26/13 on the promotion, protection and enjoyment of human rights on the Internet, as well as General Assembly resolutions 68/167 and 69/166 on the right to privacy in the digital age, to guarantee full respect for human rights, including the right to freedom of expression" (Brown, Esterhuysen, and Kumar, 2019: 5).
2. States should implement measures to detect, monitor, and mitigate disinformation and influence operations.
3. States should cooperate with information exchange on detected operations with other nations to aid global detection and monitoring.
4. States should respond to requests for assistance in mitigating disinformation and influence operations.
5. States should foster multi-stakeholder environment for addressing disinformation and influence operations.
6. States should institute media-literacy and capacity building programmes to mitigate the impact of disinformation and influence operations

7. States should not conduct or support disinformation or influence operations that will interfere with another states' national processes (e.g. elections).
8. States should not commandeer computing resources or accounts for distribution of disinformation or influence operations.
9. States should not knowingly allow their territory to be used for disinformation or influence operations, and should discourage non-state actors within their jurisdiction from participating in such activity operations.
10. States should not proliferate capabilities for conducting disinformation and influence operations.

Table 2 presents key themes from the relevant documents discussed above, and indicates which of the proposed norm they align to (the norms are presented after Table 2). As illustrated, the GGE and the Paris Call each aligned to 6 of the key themes and norms, with the GCSC aligning to 5 key themes and norms. ENISA & EEAS, together with Gleicher (2022a; 2022b) aligned to three of the key themes and norms, and Pamment (2022) to only two. The most prevalent alignment across documents is to Norm 2, which is the implementation of measure to mitigate disinformation.

Table 2: Alignment of key themes to documents and proposed norms

	Human rights	**Cooperation & partnerships**	**Capacity building**	**Assistance**	**Electoral processes**	**Implement legislation & policies**	**Refrain from malicious actions**	**Information sharing and reporting**	**Implement mitigating measures**	**Non-proliferation**	**Mitigate use of territory**	**Mitigate non-state activity**	**Commandeering of resources**
GGE	X	X		X			X	X			X		
GCSC					X	X		X	X			X	X
Paris Call		X	X		X				X	X		X	
Pamment (2022)			X			X			X				
ENISA & EEAS		X						X	X				
Gleicher (2022a; 2022b)		X				X		X	X				
Proposed norm	**1**	**3 & 5**	**6**	**4**	**7**	N/A	N/A	**3**	**2**	**10**	**9**	**9**	**8**

5. Conclusion

Disinformation and influence operations have been recognised as a growing concern for international security; however, the international discussion is not as mature as those for cyber operations which has multiple recognised international norms and principles. Discussions have emerged placing disinformation as a cybersecurity problem, and similarities between addressing disinformation and cybersecurity have been considered in research and official documents. This indicates that the concept of cyber norms for cybersecurity may be a basis for developing norms on state behaviour for addressing influence operations.

In this paper, thematic analysis is used to analyse existing cyber norms and disinformation documents in order to identify major themes related to responsible state behaviour and disinformation. Thirteen themes were identified, and from these a framework of 10 norms for responsible state behaviour for influence operations is proposed. Due to concerns over possible Human Rights abuses through the implementation of legislation, promoting human rights was favoured over the theme of legislation and policies. As with cybersecurity, multi-stakeholder and international cooperation is important, and nations should discourage non-state actors from contributing to influence operations and disinformation, and aid efforts to disrupt such operations when they are detected.

The significance of this paper is the application of the cybersecurity normative frameworks to the influence operations space, and to propose norms specific to disinformation. This further demonstrates an overlap between cybersecurity and influence operations, indicating a need for the two communities to engage. It is necessary to implement other measures to support a normative framework, such as confidence building

measures, which have been recognised in cybersecurity and international security fields. However, this is outside the scope of the current paper, and will be considered in future research.

Acknowledgements

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant no. 150381, received by the first author). The opinions, findings and conclusions or recommendations expressed in this paper are those of the authors, and not of the respective institutions or funding agencies.

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/

Allen-Ebrahimian, B. (2020) Bots boost Chinese propaganda hashtags in Italy, Axios, 1 April, [online], accessed 12 January 2023, https://www.axios.com/bots-chinese-propaganda-hashtags-italy-cf92c5a3-cdcb-4a08-b8c1-2061ca4254e2.html

Brown, D., Esterhuysen, A., and Kumar, S. (2019) Unpacking the GGE's framework on responsible state behaviour: Cyber norms, Association for Progressive Communications and Global Partners Digital, 23 December, [online], accessed 4 January 2022, https://www.apc.org/en/pubs/unpacking-gges-framework-responsible-state-behaviour-cyber-norms.

Burke, J. (2017) "US woman charged over tweet allegedly insulting Robert Mugabe", *The Guardian*, 3 November, [online], accessed 14 January 2023, https://www.theguardian.com/world/2017/nov/03/martha-odonovan-arrested-zimbabwe-alleged-mugabe-goblin-tweet-harare

Carnegie Endowment for International Peace. (2023) Partnership on Countering Influence Operations, [online], accessed 12 January 2023, https://carnegieendowment.org/specialprojects/counteringinfluenceoperations/

Crowley, K. (2022) "5 Advantages of Deep Learning in Cybersecurity", Deep Instinct, 7 September, [online], accessed 30 December 2022, https://www.deepinstinct.com/blog/five-advantages-of-deep-learning-for-preventing-cybersecurity-threats

De Wet, P. (2020) "SA expects WhatsApp to 'immediately' remove fake coronavirus news under new rules", *Business Insider*, 27 March, [online], accessed 14 January 2023, https://www.businessinsider.co.za/ott-fake-news-rules-against-coronavirus-for-whatsapp-2020-3

DFRLab. (2023) 360/Digital Sherlocks, Carnegie Endowment for International Peace, [online], accessed 12 January 2023, https://www.digitalsherlocks.org/360os-digitalsherlocks

Digital Watch (2023). UN OEWG and GGE, Geneva Internet Platform, [online], accessed 12 January 2023, https://dig.watch/processes/un-gge

DISARM Foundation. (n.d.) DISARM Framework, [online], accessed 27 December 2022, https://www.disarm.foundation/framework

ENISA and EEAS. (2022). Foreign Information Manipulation and Interference (FIMI) and Cybersecurity – Threat Landscape, 8 December, [online], accessed 28 December 2022, https://www.enisa.europa.eu/publications/foreign-information-manipulation-interference-fimi-and-cybersecurity-threat-landscape

Federal Trade Commission. (2022) *Combatting Online Harms Through Innovation*, Report to Congress, 16 June, [online], accessed 30 December 2022, https://www.ftc.gov/system/files/ftc_gov/pdf/Combatting%20Online%20Harms%20Through%20Innovation%3B%20Federal%20Trade%20Commission%20Report%20to%20Congress.pdf

Fein, D., and Stocker, E. (2022) "Cyber AI Analyst: Cutting Through the Noise to Gain the Security Edge", Darktrace, 30 November, [online], accessed 30 December 2022, https://darktrace.com/blog/cyber-ai-analyst-cutting-through-the-noise-to-gain-the-security-edge

Gleicher, N. (2022a) "Innovating against Adversaries", Keynote, *14th International Conference on Cyber Conflict: Keep Moving*, 31 May to 3 June, [online], accessed 27 December 2022, https://www.youtube.com/watch?v=ULQ4RZedRlQ

Gleicher, N. (2022b) "Innovating Against Adversaries: Exploiting the Defenders' Advantage", Lightning Talk, 360/Open Summit: Contested Realities | Connected Futures, 6-7 June, [online], accessed 27 December 2022, https://youtu.be/ZOLkgoL17dE?t=18125

Gold, H. (2020) "YouTube tries to limit spread of false 5G coronavirus claims after cellphone towers attacked", *CNN*, 6 April, [online], accessed 14 January 2023, https://edition.cnn.com/2020/04/06/tech/5g-coronavirus-conspiracy/index.html

Hodgson, T.F., Farise, K., and Mavedzenge, J. (2020) "Southern Africa has cracked down on fake news, but may have gone too far", *Mail&Guardian*, 5 April, [online], accessed 14 January 2023, https://mg.co.za/analysis/2020-04-05-southern-africa-has-cracked-down-on-fake-news-but-may-have-gone-too-far/

IEEE Computer Society. (2021) *Mitigating Societal Harms in a Social Media World*, Lessons Learned: A Reflection on the IEEE Computer Society Tech Forum, 27 September, [online], accessed 30 December 2022, https://www.computer.org/publications/tech-news/research/misinformation-disinformation-resources#report

Kahn, I. (2021) *Disinformation and freedom of opinion and expression*, Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Human Rights Council, 47th session, 21 June-9 July.

Kelley, A. (2022) "AI is 'No Magical Shortcut' FTC Says in Fighting Disinformation Online", *NextGov*, 17 June, [online], accessed 21 June 2022, https://www.nextgov.com/emerging-tech/2022/06/ai-no-magical-shortcut-ftc-says-fighting-disinformation-online/368341/

Kyriakidou, M., Cushion, S., Hughes, C., and Morani, M. (2022) "Questioning Fact-Checking in the Fight Against Disinformation: An Audience Perspective", *Journalism Practice*, DOI: 10.1080/17512786.2022.2097118

López-Marcos C., and Vicente-Fernández P. (2021) "Fact Checkers Facing Fake News and Disinformation in the Digital Age: A Comparative Analysis between Spain and United Kingdom", *Publications* 9(36), https://doi.org/10.3390/publications9030036

Lunghi, D. (2017) "From Cybercrime to Cyberpropaganda," *Trend Micro*, 16 October, [online], accessed 13 November 2017, http://blog.trendmicro.com/trendlabs-security-intelligence/from-cybercrime-to-cyberpropaganda/

Media Monitoring Africa. (2023) Real411, [online], accessed 14 January 2023, https://www.real411.co.za/

Moyo, T. (2020) "Covid-19 and the suppression of freedom of expression", *Daily Maverick*, 9 April, [online], accessed 14 January 2023, https://www.dailymaverick.co.za/article/2020-04-09-covid-19-and-the-suppression-of-freedom-of-expression-part-two/

Pamment, J. (2022) *A Capability Definition and Assessment Framework for Countering Disinformation, Information Influence, and Foreign Interference*, NATO Strategic Communications Centre of Excellence, [online], accessed 17 January 2023, https://stratcomcoe.org/publications/a-capability-definition-and-assessment-framework-for-countering-disinformation-information-influence-and-foreign-interference/255

Ramluckan, T., Wanless, A., and van Niekerk B. (2019) "Cyber-Influence Operations: A Legal Perspective", *Proceedings of the 18th European Conference on Cyber Warfare and Security* (ECCWS), Coimbra, Portugal, 4-5 July, pp. 379-388.

Ramluckan, T., and van Niekerk B. (2019) "International Humanitarian Law and Cyber-Influence Operations", *Journal of Information Warfare* 18(3), 67-82.

Roache, M, Tewa, S, Cadier, A, Labbe, C, Padovese, V, *et al*. (2022) "Russia-Ukraine disinformation tracking center: 303 websites spreading war disinformation and the top myths they publish", *NewsGuard*, 15 November, [online], accessed 16 November 2022, https://www.newsguardtech.com/special-reports/russian-disinformation-tracking-center/

Rid, T. (2020) *Active Measures: The Secret History of Disinformation and Political Warfare*, New York: Farrar, Straus and Giroux.

Schmitt, M.N. (2017) *Tallinn Manual 2.0: On the International Law Applicable to Cyber Operations*, Cambridge: Cambridge University Press.

Simon, S. (2022) "Russian law bans journalists from calling Ukraine conflict a 'war' or an 'invasion'", *NPR*, 5 March, [online], accessed 14 January 2023, https://www.npr.org/2022/03/05/1084729579/russian-law-bans-journalists-from-calling-ukraine-conflict-a-war-or-an-invasion

Stengel, R. (2019) *Information Wars: How we Lost the Global Battle Against Disinformation and what we can do about it*, New York City: Atlantic Monthly Press.

Terp, S.J., Breuer, P. (2022) "DISARM: a Framework for Analysis of Disinformation Campaigns", 2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), 6-10 June. Available at: https://ieeexplore.ieee.org/document/9830669

The Paris Call for Trust and Security in Cyberspace. (2018) The 9 Principles, 12 November, [online], accessed 5 January 2023, https://pariscall.international/en/principles

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.

United Nations General Assembly. (2021) *Countering disinformation for the promotion and protection of human rights and fundamental freedoms*, Resolution 76/277, 24 December.

Viral Facts Africa. (2023) Viral Facts Africa, Twitter, [online], accessed 14 January 2023, https://twitter.com/viralfacts

Wanless, A., and Shapiro, J.N. (2021) Why Are Authoritarians Framing International Approaches to Disinformation? *LawFare Blog*, 28 December, [online], accessed 17 February 2022, https://www.lawfareblog.com/why-are-authoritarians-framing-international-approaches-disinformation

Wanless, A., and Shapiro, J.N. (2022) *A CERN Model for Studying the Information Environment*, Carnegie Endowment for International Peace, 17 November, [online], accessed 28 December 2022, https://carnegieendowment.org/2022/11/17/cern-model-for-studying-information-environment-pub-88408

The Identification of Cybersecurity Work Roles for the Water Sector in South Africa

Aby Melanie[1], Annlize Marnewick[1] and Sune von Solms[2]
[1]Postgraduate School of Engineering Management, University of Johannesburg, South Africa
[2]Department of Electrical and Electronic Engineering, University of Johannesburg, South Africa

aby.melany@gmail.com
amarnewick@uj.ac.za
svonsolms@uj.ac.za

0000-1111-2222-3333
1111-2222-3333-4444
0000-0002-1857-1683

Abstract: This paper presents the results of a content analysis conducted on the work roles of cybersecurity practitioners for the water sector of South Africa. The paper presents literature review findings on national and international frameworks and guidelines detailing cyber security considerations for the South African water sector as well as national and international guidelines and frameworks which detail the various work roles carried out by cybersecurity practitioners in an organisation. The study found that cyber security considerations and work roles such as physical security of assets, testing and assessment of cybersecurity methods, supply chains cyber security as well as incident investigation and interfacing with law enforcement, were not well defined for cyber security discipline. The study delivers a framework detailing the work roles of cybersecurity practitioners which can be applied to the South African water sector.

Keywords: Cybersecurity, work roles, water sector

1. Introduction

The water sector has attempted to bridge the gap between operating and information technologies by overlapping these systems to reduce maintenance costs and optimise the control and monitoring systems. The potential for cyber threats has increased due to the overlapping of these systems, (Wei et al., 2010; Skiba, 2020). It is important to note that critical infrastructure, such as the water infrastructure, is vulnerable if the information technology systems are not technically well protected, (Cavelty, 2007; Germano, 2019). By not having it protected, this leads to potential cyber-attacks which is detrimental to the conservation of water quality and quantity, (Stouffer and Candell, 2014; Park and Park, 2018). The resources and capabilities for detecting, preventing, and mitigating cyber risks are currently inadequate for many utilities, (Germano, 2019). To circumvent this, the National Cybersecurity Policy Framework (NCPF) was released by the South African government in 2015, however the framework does not outline the cybersecurity practitioners work roles for the water sector of South Africa (SA), (State Security Agency, 2015). This work aims to rectify this shortcoming by studying the defined work roles within the National Institute of Standards and Technology (NIST) National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework (CWF), (Petersen *et al.*, 2020), and other international and national best practice guidelines and frameworks to develop a work roles framework for cyber security practitioners in the water sector of SA. The focus of this research was on the water sector, but can be utilised in other critical infrastructure sectors as the results prove to be overarching and applicable to the wider critical infrastructure sector. To develop the framework, the study defines the water sector organisational structure from literature, the cybersecurity considerations, and based on the defined considerations, define cybersecurity work roles and close any gaps identified. The development of this framework will assist in detection, mitigation and prevention of future cyber threats.

2. Literature Review

2.1 Water sector organisational structure

To start to address the concerns around the threat to water quantity and quality, the South African water sector organistional structure needed to be understood. Legislation and policies is developed at the ministerial level, (Beck *et al.*, 2016). Departmental level is responsible for governance related to policies and legislation. This level also ensures and enforces compliance, (Ruiters and Matji, 2015). National and regional levels is responsible for the development and management of all procedures, (GreenCape, 2014). Local level is responsible to ensure

that policies and procedures are carried out within their jurisdictions. Plant level carries out and implements all procedure and policies. This level will also monitor, control and report any anomalies which may adversely impact plant operations, (GreenCape, 2014; Ruiters and Matji, 2015; Beck *et al.*, 2016).

2.2 Frameworks and guidelines to define the cybersecurity considerations for the water sector

Internationally, a variety of cybersecurity frameworks, guidelines and standards have been produced or are in the process of being developed, (Panguluri, Phillips and Cusimano, 2011; Germano, 2019). The International Organisation of Standards (ISO) 27002 has been adopted by South African Bureau of Standards (SABS) which specifies an organisation's minimum information security criteria, (International Organisation for Standardisation, 2013). Other data sources identified includes:

- Guide to Industrial Control Systems (ICS) Security, (Stouffer et al., 2015),
- 15 cybersecurity fundamentals for water and wastewater utilities: Best practices to reduce exploitable weaknesses and attacks, (Water Information Sharing and Analysis Center, 2019)
- American Water Works Association (AWWA) Water sector cybersecurity risk management guidance, (Yost, 2019)
- Roadmap to secure control systems in the water sector, (Water Sector Coordinating Council Cyber Security Working Group, 2008).

The documents mentioned can be used to define the cyber security considerations for the water sector of South Africa.

Gaps identified from literature indicates that the cyber security industry is relatively new and many of its standards and guidelines are still in development, (Panguluri *et al.*, 2011). Not all the cyber security requirements for the water sector have been identified but rather has been developed based on known threats and vulnerabilities, (Water Sector Coordinating Council Cyber Security Working Group, 2008; Germano, 2019; Yost, 2019). Literature also indicates that methods of physical security, access and authentication, software enhancements and privacy enhancements are all cyber security components which are lacking. Another gap identified is the requirements for improving the security between the business and ICS network as well as intrusion detection, (Panguluri *et al.*, 2011; Clark *et al.*, 2016; Germano, 2019).

2.3 Frameworks and guidelines to define the work roles of cybersecurity practitioners

Skills Framework for the Information Age (SFIA) is a model for characterising and managing skills and competencies for Information and Communication Technologies (ICT) and cyber security professionals. The model assists with understanding IT skills in general and it contains security-related skills, (Furnell, 2021). SFIA describes the abilities and competencies required by professionals in positions involving information and communication technology and cyber security, (SFIA Foundation, 2018).

According to Caulkins, Marlowe and Reardon, (2019), the demand for cyber security practitioners is growing. Competent cybersecurity practitioners are hard to come by in all industries and companies, (Campbell, O'Rourke and Bunting, 2015). The NIST NICE CWF acts as a guiding principle for the development of a unified cyber security workforce. The NIST NICE CWF details the requirements for work force identification, standardises development of work role descriptions, qualification requirements and training requirements for the development of a capable and ready workforce, (Dawson, Taveras and Taylor, 2019; Schmeelk and Dragos, 2020). The NIST NICE CWF was released in order to develop a process for determining specific cybersecurity work roles (Campbell *et al.*, 2015). The NIST NICE CWF was used as the baseline data source for the study.

Gaps identified shows that the educational programs at various institutes are not aligned to the NIST NICE CWF, especially in the Analyse and Investigate categories, (Caulkins *et al.*, 2019; Saharinen, Backlund and Nevala, 2020). Another gap is that the NIST NICE CWF focusses on the technical skills required for the cyber security work force however, research, (Campbell *et al.*, 2015; Caulkins *et al.*, 2019), indicates that non-technical skills are just as important as that of the technical kind. Based on the current professionals within the industry, the work roles framework that was identified may not be implementable due to the lack of educational programs available, grasp of non-technical skills and the availability of experienced individuals.

3. Methodology

This study adopted the qualitative content analysis methodology using secondary data identified from literature. This methodology is utilised because it assures that the original data is preserved, the data is analysed systematically rather than selectively, and it is used to analyse qualitative data and construct frameworks, (Lancaster, 2005; Sekaran and Bougie, 2016). Firstly, the cybersecurity considerations for the water sector were defined by using the ISO 27002 as the baseline data source. Verification and validation of the considerations extracted from ISO 27002 was performed by using the other data sources identified in section 2.2. Secondly, the work roles for cyber security practitioners were defined by listing all the cyber security work roles and their definitions from the NIST NICE CWF. A meaning unit was defined, in the case of this research, defining the meaning unit refers to condensing of the work roles definition. The meaning unit was then categorised and coded. The code was used to match the work role to a specific cyber security consideration. Validation and verification were carried out by using SFIA. The results from the data analysis process were interpreted to construct a framework to address the research aim.

4. Results, Analysis and Discussions

4.1 Water sector cybersecurity considerations

To establish the appropriate work roles for cybersecurity practitioner in the water sector, the cybersecurity considerations for the South African water sector was firstly determined. The considerations were collected from four (4) different data sources, as shown in sections 2.2. The data collected was combined and examined to create a comprehensive set of cybersecurity consideration for the water sector in SA. Firstly, ISO 27002, was used to list and define the baseline of cybersecurity considerations. Secondly, these cybersecurity considerations were verified against the three (3) other data sources. They were verified by performing latent and manifest analysis on the definitions of the considerations. A total of fourteen (14) considerations was defined.

4.2 Cybersecurity practitioners work roles for the South African water sector

Data was collected from the NIST NICE CWF and analysed to define the work roles of cyber security practitioners in the water sector based on cyber security considerations defined in section 4.1. SFIA was used for verification and validation. **Fig. 1** shows the steps that were followed to develop a comprehensive set of cyber security practitioner work roles for the water sector in SA.

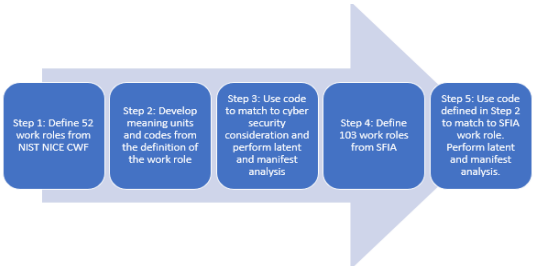


Fig. 1. Steps followed to develop a comprehensive set of cyber security practitioner work roles for the water sector in SA.

Fig. 2 illustrates how steps 1-3 was applied to the data analysis process by using the work role *Risk Management RSK: Authorising Official* from the NIST NICE CWF.

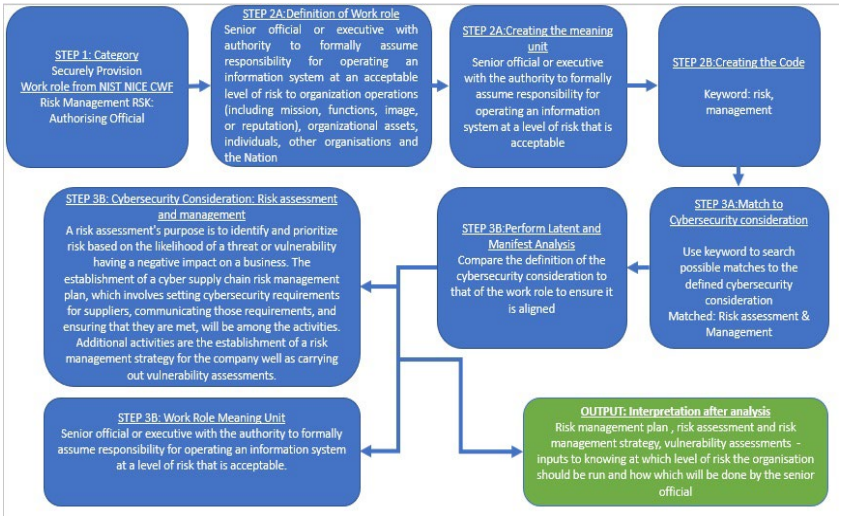


Fig. 2. Implementation of step 1-3 on one work role identified from NIST NICE CWF

Fig. 3 illustrates the application of the verification and validations steps 4- 5 on the work role, *Risk Management RSK: Authorising Official* from the NIST NICE CWF.

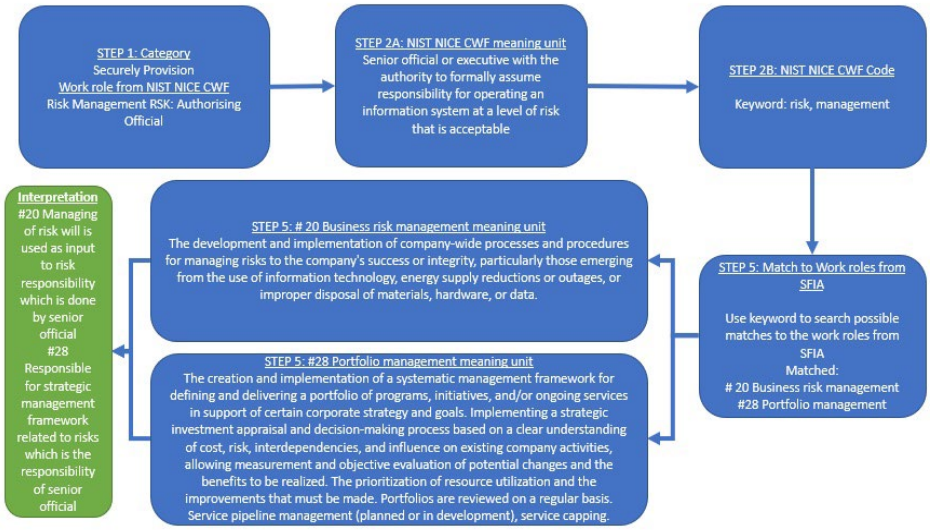


Fig. 3. Implementation of steps 4-5 on one work roles identified form NIST NICE CWF

Table 1 shows the summary of the results of the data analysis process for the NIST NICE CWF Work role: *Risk management RSK: Authorising Official*. The work role matched the cyber security consideration *Risk Assessment & Management*. It was also verified and validated by matching to the SFIA work roles *Business Manager* and *Portfolio Manager*. This will be added to the construction of the work roles framework.

Table 1. Results for the Risk management RSK: Authorising Official

Category	NIST NICE CWF Work role	Cyber Security Consideration	SFIA Work role
Securely Provision	Risk Management RSK: Authorising Official	Risk assessment and management	1. Business risk manager 2. Portfolio manager

By following the steps in **Fig. 1**, a list of thirty-seven (37) work roles was identified. Additionally, six (6) gaps have also been identified. Four (4) gaps were related the cyber security considerations and two (2) gaps was related to the defined work roles within the NIST NICE CWF.

4.3 Gaps identified

Two types of gaps have been identified. The first set of gaps were related to the water sector cybersecurity considerations. No considerations match the following work roles namely, 1. *Test & Evaluation: System Test & Evaluation Specialist*, 2. *Cyber Investigation: Cyber Crime Investigator*, 3. *Digital Forensics: Law Enforcement/Counterintelligence Forensics Analyst* and 4. *Digital Forensics: Cyber Defence Forensics Analyst.* The second set of gaps were related to the NICE NIST CWF work role. The work roles did not match the following cyber security considerations: 1. *Physical & Environmental Security* and 2. *Securing the Supply Chain*. The gaps related to the cybersecurity considerations revealed that little emphasis is currently upon identifying where the cyber threats emanate from, as well as apprehension of the perpetrators. This is a very important aspect to consider as this will influence consequences applied to cyber attackers and future cyber-criminal activity. Gaps related to the defined work roles indicate that third party cyber-attacks are currently not considered as a high risk. Cyber-attacks can come from any source and may have devastating impacts on an organisation. Another gap identified was related to physical security of cybersecurity assets. Hardware can easily be infiltrated by a cyber attacker. To close the gap related to the cybersecurity consideration and the defined work roles from NIST NICE CWF, the following was done:

- The definitions of the required consideration was adapted from the definition of the work role identified as well as the definition from SFIA.
- The methodology of creating new work roles was applied by defining the task by adapting a definition from SFIA and the cybersecurity considerations.

4.4 Framework for cybersecurity practitioners for the South African water sector

The results from the data analysis process were used to develop a framework for cyber security practitioners which defines the cyber security work roles for the water sector of SA. The gaps identified was addressed and new work roles and cybers security considerations were defined. The now complete list of work roles and cyber security considerations were matched to the levels within the water sector organisation structure which was defined from literature.

Table 2. Cyber security work roles for the water sector of SA Framework

Cyber security considerations	Work Role	Organisation structure
Governance	Executive Cyber Leadership: Executive Cyber Leadership	Ministerial
	Strategic Planning and Policy: Cyber Policy and Strategy Planner	Departmental
Human Resource Security & Cyber security Awareness	Strategic Planning and Policy: Cyber Workforce Developer and Manager	Departmental
Compliance	Legal Advice and Advocacy: Cyber Legal Advisor	National & Regional
	Legal Advice and Advocacy: Privacy Officer/Privacy Compliance Manager	National & Regional
Risk assessment & Management	Risk Management (RSK): Authorising Official	National & Regional
Governance	Conducts evaluations of an IT program or its individual components, to determine compliance with published standards	Local
Business Continuity, incidents, emergencies and Disaster Recovery planning	Cyber Operational Planning: Cyber Intel Planner	Local
	Cyber Operational Planning: Cyber Ops Planner	Local
Cryptography, Design and implementation of improved security systems	Systems Development: Systems Developer	Local
	Systems Architecture: Security Architect	Local
Participate in Partnership and Outreach for information sharing and collaboration	Cyber Operational Planning: Partner Integration Planner	Local
	Cyber Defence Infrastructure Support: Cyber Defence Infrastructure Support Specialist	Local
Human Resource Security & Cyber security Awareness	Training, Education and Awareness: Cyber Instructional Curriculum Developer	Local
	Training, Education and Awareness: Cyber Instructor	Local
Forensic analysis and law enforcement liaison for cyber crimes	Digital Forensics: Law Enforcement/Counterintelligence Forensics Analyst	Local
	Digital Forensics: Cyber Defence Forensics Analyst	Local
Governance	Cyber Security Management: Information Systems Security Manager	Works

Cyber security considerations	Work Role	Organisation structure
Business Continuity, incidents, emergencies and Disaster Recovery planning	Incident Response: Cyber Defense Incident Responder	Works
Risk assessment & Management	Risk Management (RSK): Security Control Assessor	Works
Cryptography, Design and implementation of improved security systems	Software Development: Software Developer	Works
	Software Development: Secure Software Assessor	Works
Participate in Partnership and Outreach for information sharing and collaboration	Technology R&D: Research and Development Specialist	Works
Cyber security Testing & Evaluation	Test & Evaluation: System Test & Evaluation Specialist	Works
Cryptography, Design and implementation of improved security systems	Systems Development: Information Systems Security Developer	Works
Asset Management	Data Administration: Database Administrator	Works
	Data Administration: Data Analyst	Works
	Knowledge Management: Knowledge Manager	Works
Access Control	Systems Analysis: Systems Security Analyst	Works
Communication Security	Cyber Security Management: Communications Security Manager	Works
Operations Security	Cyber Defence Analysis: Cyber Defence Analyst	Works
Systems acquisition, development and maintenance	Vulnerability Assessment and Management: Vulnerability Assessment Analyst	Works
	Warning/Threat Analysis: Threat/Warning Analyst	Works
	Exploitation Analysis: Exploitation Analyst	Works
	All-Source Analysis: All-Source Analyst	Works
	All-Source Analysis: Mission Assessment Specialist	Works
	Targets: Target Developer	Works
	Targets: Target Network Analyst	Works
	Language Analysis: Multi-Disciplined Language Analyst	Works
Operations Security	Cyber Operations: Cyber Operator	Works
Investigation of cyber crimes	Cyber Investigation: Cyber Crime Investigator	Works
Physical & Environmental Security	Security Officer	Works
Securing the Supply Chain	Third Party Cyber Security Officer	Works

Table 2, shows all Forty-three (43) work roles which have been defined to a level within the water sector's organisational structure. It should be noted that the framework indicates the minimum work roles required based on the water sector organisational structure developed from literature.

5. Recommendations and Conclusion

The work roles defined for cyber security practitioners should be filled by the organisation to ensure that prevention, mitigation and detection of cyber threats occur to reduce this emerging risk. The water sector needs to start sharing experiences of cyber threats and events across sectors as this will assist in implementing lessons learnt and possibly preventing future incidents. Application of the framework needs to start at ministerial level where legislation must be developed, reviewed and implemented and a clear plan for implementation is also developed. Departmental level of the water sector ensures that the legislation, policies and procedures are governed and enforced by measuring compliance. This level must also ensure that resources is made available to carry out the required cyber security practitioner work roles. National and regional levels of the water sector must ensure the development of procedures and management of thereof. Local level must ensure that policies and procedures are carried out correctly within their jurisdiction. Plant level must ensure that legislation, policies and procedure are carried out. The findings of this study should be used to identify existing resources to meet the cyber security workforce needs. Secondly, an assessment of the educational programs needs to be done to ensure alignment to the required work roles for the sector.

References

Beck, T., Rodina, L., Luker, E. and Harris, L. (2016) *Institutional and Policy Mapping of the water sector in South Africa*. Cape Town. doi: 10.13140/RG.2.2.32761.88164.

Campbell, S. G., O'Rourke, P. and Bunting, M. F. (2015) 'Identifying dimensions of cyber aptitude: The design of the cyber aptitude and talent assessment', *Proceedings of the Human Factors and Ergonomics Society*, 2015-Janua, pp. 721–725. doi: 10.1177/1541931215591170.

Caulkins, B., Marlowe, T. and Reardon, A. (2019) 'Cybersecurity Skills to Address Today's Threats', *Advances in Intelligent Systems and Computing*, 782, pp. 187–192. doi: 10.1007/978-3-319-94782-2_18.

Cavelty, M. D. (2007) *Critical information infrastructure_ vulnerabilities, threats and responses*. 1st edn, *ICTs and International Security*. 1st edn. Zurich, Switzerland.

Clark, R. M., Panguluri, S., Nelson, T. D. and Wyman, R. P. (2016) *Protecting Drinking Water Utilities from Cyber Threats*. Idaho.

Dawson, M., Taveras, P. and Taylor, D. (2019) 'Applying Software Assurance and Cybersecurity NICE Job Tasks through Secure Software Engineering Labs', *Procedia Computer Science*, 164, pp. 301–312. doi: 10.1016/j.procs.2019.12.187.

Furnell, S. (2021) 'The cybersecurity workforce and skills', *Computers and Security*, 100, p. 102080. doi: 10.1016/j.cose.2020.102080.

Germano, J. H. (2019) *Cybersecurity Risk & Responsibility in the Water Sector*. New York.

GreenCape (2014) *Market Intelligence Report: Water*. Cape Town.

International Organisation for Standardisation (2013) *ISO/IEC 27002:2013 Information Technology- Security techniques - Code of practice for information security controls*. ISO/IEC.

Lancaster, G. (2005) *Research Methods in Management: A Concise introduction to research in management and business consultancy*. Burlington, MA: Elsevier Butterworth-Heinemann.

Panguluri, S., Phillips, W. and Cusimano, J. (2011) 'Protecting water and wastewater infrastructure from cyber attacks', *Frontiers of Earth Science*, 5(4), pp. 406–413. doi: 10.1007/s11707-011-0199-5.

Park, S.-N. and Park, D.-W. (2018) 'Cybersecurity System for Water Treatment SCADA System', *Journal of Engineering and Applied Sciences*, 13(11), pp. 8712–8715. Available at: http://library1.nida.ac.th/termpaper6/sd/2554/19755.pdf.

Petersen, R., Santos, D., Smith, M. C., Wetzel, K. A. and Witte, G. (2020) *Workforce Framework for Cybersecurity (NICE Framework)*. Gaithersburg, MD. doi: 10.6028/NIST.SP.800-181r1.

Ruiters, C. and Matji, M. P. (2015) 'Water institutions and governance models for the funding, financing and management of water infrastructure in South Africa', *Water SA*, 41(5), pp. 660–676. doi: 10.4314/wsa.v41i5.09.

Saharinen, K., Backlund, J. and Nevala, J. (2020) 'Assessing Cyber Security Education through NICE Cybersecurity Workforce Framework', *ACM International Conference Proceeding Series*, pp. 172–176. doi: 10.1145/3436756.3437041.

Schmeelk, S. and Dragos, D. (2020) 'Wireless Security: Examining the next NICE Framework Iteration based on Industry Requirements', *Cybersecurity Skills Journal: Practice and Research*, (Special), pp. 59–73.

Sekaran, U. and Bougie, R. (2016) *Research Methods for Business: A skills building approach*. Seventh Ed. John Wiley & Sons Ltd.

SFIA Foundation (2018) *Skills Framework of the Information Age (SFIA): The complete reference*. England. Available at: https://www.sfia-online.org/en.

Skiba, R. (2020) 'Water Industry Cyber Security Human Resources and Training Needs', *International Journal of Engineering Management*, 4(1), p. 11. doi: 10.11648/j.ijem.20200401.12.

State Security Agency (2015) *The National Cybersecurity Policy Framework for South Africa*. Available at: www.gpwonline.co.za.

Stouffer, K. and Candell, R. (2014) 'Measuring impact of cybersecurity: On the performance of industrial control systems', *Mechanical Engineering*, 136(12), pp. 4–7. doi: 10.1115/1.2014-dec-5.

Water Information Sharing and Analysis Center (2019) *15 Cybersecurity Fundamentals for Water and Wastewater Utilities: Best practices to reduce exploitable weaknesses and attacks*. Washington. Available at: www.waterisac.org.

Water Sector Coordinating Council Cyber Security Working Group (2008) *Roadmap Secure Control Systems in the water sector*, *Control*. Washington DC.

Wei, D., Lu, Y., Jafari, M., Skare, P. and Rohde, K. (2010) 'An integrated security system of protecting smart grid against cyber attacks', in *Innovative Smart Grid Technologies Conference, ISGT 2010*. doi: 10.1109/ISGT.2010.5434767.

Yost, W. (2019) *Water Sector Cybersecurity Risk Management Guidance*.

Cyber Lessons that the World Can Learn from Lithuania

M.Warren[1], D.Štitilis[2] and M.Laurinaitis[2]
[1]RMIT University, Melbourne, Australia and Univeristy of Johannesburg, Johannesburg, South Africa
[2]Mykolas Romeris University, Vilnius, Lithuania.

Matthew.warren2@rmit.edu.au
stitilis@mruni.eu
laurinaitis@mruni.eu

Abstract: In an era of rapid technological advancements and increasing online connectivity, the proliferation of cyber threats, including the spread of fake news and disinformation, presents a significant challenge to nations worldwide. Lithuania has emerged as a leading example in addressing these challenges, particularly concerning cyber groups such as Killnet and disinformation / fake news. This paper aims to explore the key cyber lessons that can be learned from Lithuania's proactive approach in dealing with Killnet and combating disinformation / fake news. By analysing Lithuania's cybersecurity strategies and initiatives, this paper identifies crucial lessons that can be applied globally. Firstly, Lithuania recognises the importance of co-ordinated cyber security technologies and national frameworks to counter cyber groups such as Killnet attacks. Secondly, Lithuania has effectively tackled the spread of fake news / disinformation through a comprehensive approach involving legislation, media literacy programs, and strong cooperation between government agencies, civil society organisations, and the private sector. The country's experience underscores the significance of collaborative efforts in combating misinformation, promoting media literacy, and fostering critical thinking skills among the population.

Keywords: Cyber Security, Lithuania, Russia, Fake News and KillNet.

1. Introduction and Background

Lithuania is a unique country when it comes to cyber security due to its whole-of-government and whole-of-society approach to the issue. The country has adopted a national cyber security strategy that focuses on developing cyber defence capabilities, building a secure state data-transfer network, and promoting innovative cyber security tools. Lithuania is also recognised as one of the best-prepared countries in the world for cyber security and leads the EU's (European Union) cyber security cooperation efforts. Moreover, the country emphasises public awareness and education to deal with evolving threats. Lithuania's unique approach to cyber security emphasises cooperation and education to deal with the evolving threats of cyber security attacks and fake news (disinformation) campaigns.

Lithuania is a country located in the Baltic, with a population of 2.8 million, a member of the European Union, and NATO and Lithuania is considered a front-line cyber state. According to the International Telecommunication Union - Global Cyber Security Index, Lithuania is ranked 6th globally (ITU, 2021) for their national cyber security capability and resilience.

Historically Lithuania has developed a unique national approach to cyber security. The Lithuanian Cyber Security Law, implemented in 2015, was one of the first laws of its kind in Europe, defined critical infrastructure in Lithuania and established security requirements for Lithuanian critical infrastructure providers (Baltic News, 2018).

The first Lithuanian National Cyber Security Strategy was developed by the Lithuanian Ministry of National Defence in collaboration with other government agencies and industry stakeholders in 2017 (Ministry of National Defence, 2017). Key principles of the Lithuanian National Cyber Security strategies include (Ministry of National Defence, 2017, Ministry of National Defence, 2019):

- strengthening legal and regulatory frameworks: Lithuania has enacted legislation and established regulatory frameworks to address cyber threats effectively. This includes the Law on Cybersecurity, which provides a legal basis for preventing, detecting, and responding to cyber incidents;
- building cybersecurity capabilities: The strategy emphasizes the development of cybersecurity expertise and skills among professionals, law enforcement agencies, and the public. Lithuania has invested in cybersecurity education, training programs, and international cooperation to enhance its capabilities;
- promoting public-private partnerships: Recognising the importance of collaboration, Lithuania actively fosters partnerships between government institutions, private sector entities, and academia. This approach encourages information sharing, joint initiatives, and coordinated responses to cyber threats;

- enhancing international cooperation: Lithuania actively engages in international efforts to combat cybercrime and improve global cybersecurity.

The updated 2019 Lithuanian National Cyber Security Strategy outlines a range of objectives for Lithuania, including enhancing the country's cyber security capabilities, developing a strong legal and regulatory framework for cyber security and promoting international cooperation in the cyber domain (Ministry of National Defence, 2019). Lithuanian has also set up the Regional Cyber Defence Centre (RCDC) which is a subsidiary of the National Cyber Security Centre under the Ministry of National Defence of the Republic of Lithuania. The key operational objectives of the RCDC (RCDC, ND are):

- to strengthen cooperation with the United States of America, Ukraine, Georgia, Poland and other strategic partners of Lithuania in the field of cyber security;
- to carry out cyber threat analysis jointly with the partners;
- to organise training courses for cyber security specialists;
- to conduct international scientific research in the area of cyber security.

As cited in Warren 2023 (Warren et al, 2023) Russia has a long history of using information and cyber security as a weapon in offensive and defensive manners, from Estonia 2007, Georgia 2008, Ukraine 2014, 2022 as well as numerous examples of election interference in Western countries. Ukraine has been dealing with Russia as an adversary since the first Russian-Ukraine war in 2014, the illegal annexation of Crimea in 2014, and the Russian invasion in 2022 (Czosseck, 2011, Bowen, 2021).

On 24 February 2022, Russia invaded Ukraine, following the invasion the Ukrainian Government had been supported with political, economic and military support from numerous governments around the world (including Lithuania) (BBC, 2023). Russia has also been supported, but this support has come from hacking groups, at the forefront of this support have been such groups such as KillNet. KillNet is a pro-Russian hacktivist group active since at least January 2022 and known for its DDoS (Distributed Denial of Service) attacks against countries supporting Ukraine, especially NATO countries since the Russia-Ukraine war broke out in 2022. DDoS is the primary type of cyber-attack approach employed by the group which can cause thousands of connection requests and packets to be sent to the target server or website per minute, slowing down or even stopping vulnerable systems. While KillNet's DDoS attacks usually do not cause major damage, they can cause service outages lasting several hours or even days (US Government, 2023).

According to KillNet, their funding has come from a variety of different sources including (Thales, 2023):

- private charities;
- funding directly from the group's activities;
- large crowdfunding campaigns, the donations are made in cryptocurrencies or by direct donation through applications such as the Telegram message application;
- partially self-funded through crypto mining activities;
- indirect funding from Russian intelligence services.

Figure 1 shows the KillNet group providing information via the Telegram application for crypto donations via Bitcoin, Ethereum, Tether payment systems

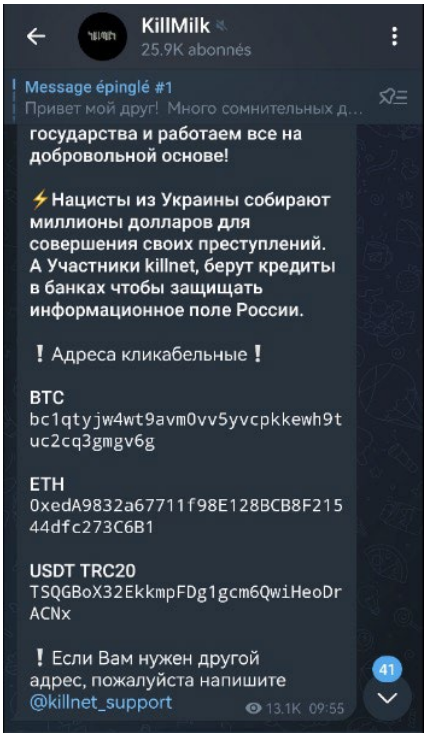


Figure 1: KillNet request for donations via Telegram application (information in Russian).

During the current Russian situation, Western organisations such as Microsoft are undertaking extensive analysis of Russian tactics and approaches. Microsoft has identified that some of the most common methods used against Western government, critical infrastructure and corporate systems include (Microsoft 2022a, Microsoft 2022b and Microsoft 2022c):

- Advanced Persistent Threats (APT): This refers to long-term, targeted attacks that are carried out by state-sponsored hacking groups. APT attacks typically involve the use of sophisticated malware and social engineering tactics to gain access to a target's network and steal sensitive information;
- phishing: This method involves tricking victims into providing login credentials or other personal information through fake emails, social media messages, or other forms of communication;
- Distributed Denial of Service (DDoS) attacks: This method involves overwhelming a website or network with a flood of traffic, making it difficult or impossible for legitimate users to access it;
- watering hole attacks: This method involves compromising a specific website that is likely to be visited by targeted victims, and using that website to deliver malware or steal information;
- supply chain attacks: This method involves compromising a third-party vendor or service provider that has access to a target's network, in order to gain access to the target's network.

Another major issue facing countries around the world is the impact of fake news also known as disinformation. What is fake news, according to Warren (2019) there are numerous aspects including:

- fabricated content: completely false content;
- manipulated content: distortion of genuine information or imagery, for example, a headline that is made more sensationalist, often popularised by ‘clickbait’;
- imposter content: impersonation of genuine sources, for example by using the branding of an established news agency;
- misleading content: misleading use of information, for example by presenting a comment as fact;
- false context of connection: factually accurate content that is shared with false contextual information, for example when a headline of an article does not reflect the content;
- satire and parody: presenting humorous but false stories as if they are true. Although not usually categorised as fake news, this may unintentionally fool readers.

2. Research Question

The paper will explore the unique challenges that Lithuania has experienced and highlight what other countries can learn from these real-life cyber experiences.

The key focus of the paper is:

How has Lithuania reacted to different cyber incidents during the Ukrainian situation?

In terms of this study, it was decided just to focus on Russia's cyber-related activities and just to focus broadly on cyber activities against Lithuania.

One of the issues with researching a live conflict is that it is unfolding, therefore a key issue is around collecting and using real-life research data that is valid in an emerging environment.

In terms of this study, it was decided to draw upon real-life public data drawn from the Lithuanian government cyber security assessments.

3. Case studies

The paper will focus on several real-life examples, in particular the hacking group KillNet and the impact of fake news. These will be discussed in the next section of the paper.

3.1 Example 1: Hacking Group KillNet

This case study is drawing upon real-life data from the Lithuanian Regional Cyber Defence Assessments (RCDC 2022a, 2022b, 2022c). The data identified the following incidents relating to KillNet and Lithuania and related to the first three quarters of 2022. The incidents identified in this period were:

Incident 1 – 16th May, 2022 "KillNet" and "Legion" Hacker groups Declare Cyber War against 10 Countries (numerous countries including Lithuania).

The pro-Russian hacker group "KillNet", together with volunteer Russian hackers "Legion", declared war on ten countries, including the UK, the US, Germany, Italy, Latvia, Romania, Lithuania, Estonia, Poland, and Ukraine. These countries were chosen "because of their support for Nazis and Russophobes," a popular narrative used by the Russian media to describe anyone supporting Ukraine's resistance to the Russian invasion. This attack was a simple call for action against Lithuania.

Incident 2 – 20th June, 2022 "KillNet" targets various network infrastructures in Lithuania.

The Russian hacktivist group "KillNet '' in their telegram channel posted a call for other hacking groups to help them cripple Lithuania's national network infrastructure. Therefore, they did by damaging the Lithuanian police website policija.lrv.lt. The hacktivist group also tried disrupting one of Lithuania's mobile providers but they were not successful. The attack was minimal in its success apart from a Lithuanian government web-site being disrupted for a short period.

Incident 3 – 27th June, 2002, a Russian hacking group takes credit for a wide-ranging cyber-attack on Lithuania.

The Lithuanian Governmental Tax Inspectorate, the Migration Department, Ministries and a number of other state agencies were among the targets of the KillNet hackers, according to a statement from the nation's defence minister and National Cyber Security Centre. On its Telegram channel, the Russian hacker organisation "KillNet" announced the attack, which was first directed at a Lithuanian and a Latvian online accounting system. The attack was minimal in its success.

Incident 4 - July 4, 2022 Lithuanian website was defaced.

One of the hacktivist groups associated with "KillNet", "NoName057(16)", posted on their Telegram channel about the defacement of the Lithuanian logistics company "ExpressTrip". On the company's defaced web page, there were claims about the "correct" attitude to the ban on the transit of Russian cargo to the Kaliningrad region. In addition, the defaced web page supported the Russian war against Ukraine and called it "the special operation of the Russian Federation armed forces". The attack was minimal in its success apart from a Lithuania corporate web-site being disrupted for a short period.

The four identified examples in relation to KillNet, identified a number of issues, namely:

1. KillNet overstated/overemphasised the rationale and impact of their cyber attacks (incidents 1, 2, 3 and 4);
2. The KillNet attacks were not sophisticated, e.g. web site defacement (incidents 2 and 4);
3. The messaging point-to-point app – Telegram was used to communicate the incidents (incidents 2,3 and 4), the rationale being that this platform cannot be censored or removed.

The KillNet examples highlight a number of cyber incidents, but the impact of these incidents was minor and when the cyber attacks actually took place they took the form of minor website defacements. Because of the cyber security maturity of the Lithuanian government and organisations they were able to identify and resolve the issues correctly. What the incidents did highlight was the limited ability of KillNet to disrupt Lithuanian core services and Lithuanian was able to react to these attacks.

3.2 Case 2 Study - Fake News and Lithuania

Lithuania has historically been a victim of fake news and disinformation and this situation has intensified recently since 2022 with Lithuania's support of Ukraine.

From June 2022, Lithuania imposed restrictions on the transit of sanctioned goods from Russia to its Kaliningrad exclave. The Russian response was to call the situation the "Kaliningrad Blockade" with claims that the Lithuanian government had dismantled the railway line running through Lithuanian connecting to Kaliningrad (LRT, 2022). As shown by Figure 2, many of the fake news stories are focused on an internal Russian audience and focus on reinforcing the narrative of the Russian Government regarding a blockade of Kaliningrad and the false message that the Lithuanian government was physically disrupting railway traffic between Russia and Kaliningrad.

Figure 2: Russian Fake News Story about the destruction of train tracks within Lithuania (StopFake, 2022).

The fake news campaign against Lithuania ran for a series of weeks with a focus on Russian news outlets for an internal Russian audience and then faded away as other political matters came to the forefront.

One of the unique aspects of Lithuania is how civil society is involved in dealing with the impact of fake news. Lithuanian society has reacted to the fake news situation by creating citizen-based groups called Elves, the

rationale being that elves fight trolls (Time, 2022). Based on the concept of elves there are a number of citizen organisations in Lithuania that have organically developed to combat fake news and disinformation the most notable being an organisation called Debunk.

Debunk is a citizen-based movement that started in Lithuania but has expanded to other Baltic countries, Poland, Georgia and Montenegro, as well as in the United States and North Macedonia. Debunk is an independent technology think tank and non-governmental organisation that analysis fake news and runs educational media literacy programs. Debunk has the following focused capabilities and characteristics (Debunk, NDa):

- a team of skilled analysts with backgrounds ranging from political science and history to business and media;
- national institutions in our partner countries, providing valuable insights on the situation in their respective regions;
- IT professional with a broad knowledge of AI tools which helps Debunk make the fact-checking process faster and fool-proof;
- a Lithuanian community of volunteer fact-checkers also known as elves (because of their notorious skills of hunting online trolls) to assist Debunk with their activities.

The Debunk also undertake the following key activities (Debunk, NDa):

- development of detailed disinformation reports in relation to current issues and country-linked campaigns;
- ongoing influence operations and monitoring of elections for foreign interference;
- training and certification of citizens involved in monitoring fake news.

The concept of “elves” was developed in Lithuania in 2014, following the events in Maidan Square in Ukraine that included clashes between protesters and pro-government forces of the Russian-backed president Viktor Yanukovych. Shortly after, Russia attacked Ukraine and occupied part of its territory and at the same time organised sophisticated fake news campaigns against Ukraine (Debunk, NDb).

An example of the work of Debunk is shown in Figure 3, this shows an analysis of a fake news story in 2021 regarding Lithuanian “perceived jealousy” of the Belarus government and their development of a nuclear power station with Russian support and that Lithuania pressured the Ukraine government not to use any power supplied by the nuclear reactor in Belarus.

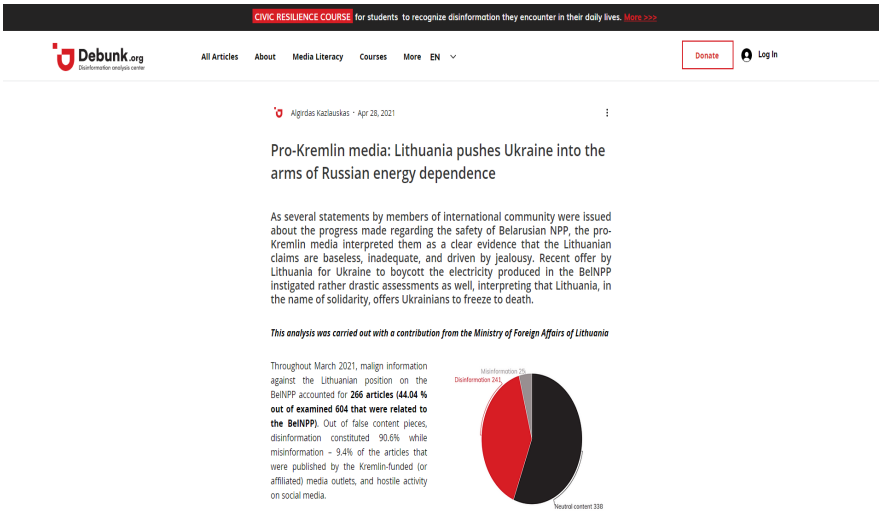


Figure 3: Debunk analysis of a Fake News story.

4. Discussion

Lithuania has been targeted by a number of Russian cyber-attacks since the Ukraine conflict started especially in terms of national DDoS attacks on Western countries. The data (see Table 1) shows that Lithuania was the 6th most attacked Western country in terms of DDoS attacks, but with Latvia in 2nd place, it shows that smaller Baltic Countries are being unproportionally targeted by these attacks.

Poland	110
Latvia	74
Sweden	60
USA	67
Germany	52
Lithuania	45

Table 1: National DDoS Attacks during the Ukraine Conflict (Thales, 2023).

In the context of this paper, many countries can learn from Lithuania's experience and expertise. In terms of the examples described in the paper and how Lithuanians responded to cyber incidents, namely:

KillNet Example

KillNet is a unique organisation, well-resourced and they have the ability to undertake cyber attacks especially DDoS attacks against a number of Western targets. Lithuania has experienced cyber security attacks against several Lithuanian government and commercial sites.

The paper reflected on four cyber incidents in 2022 related to Lithuania and KillNet and drew upon the analysis of the Lithuanian Regional Cyber Defence Centre and highlighted how Lithuania was able to mitigate these attacks and these attacks had only minimal impact.

Fake news Example

Lithuania is an ongoing target of fake news and disinformation campaigns many of which originate from Russia. Lithuania has reacted to the situation by the development of citizen initiatives that identify fake news stories and publicly disclose them as being fake news. In the age of social media and fast-spreading information, Lithuanian society is at the front line of confrontation and negative influences. Forming a strong, resilient, and critically thinking society that remains attentive to information and resistant to provocations is a key task for Lithuania, especially in the face of today's ongoing security issues (Bankauskaite and Šlekys, 2023).

There is always of risk of volunteer civic society groups such as Debunk that over time due to a lack of funding and that volunteers may leave the group due to volunteer fatigue / family commitments that these groups are not sustainable over a long period of time but now they are fulfilling a key function for Lithuanian society.

5. Conclusion

In terms of the paper, one of the challenges in reflecting upon a current conflict is the changing situation and use of appropriate data sources, this was the reason for drawing upon the data provided by the Lithuanian Regional Cyber Defence Centre and using informed data from other sources, e.g. Microsoft and Thales.

The paper has described real-life cyber challenges that are impacting the whole of Lithuania and the steps that they are undertaking to deal with the challenges. The shared Lithuania experiences is something that countries around the world can learn from. As time progress, more and more information about the current cyber and physical conflict between Russia and Ukraine will become public as more insights will be gained about groups such as KillNet and fake news campaigns against countries such as Lithuania.

By studying Lithuania's experiences in countering the Killnet group and disinformation / fake news, policymakers, cybersecurity professionals, and researchers worldwide can gain valuable insights into effective strategies, legislation, and collaborative approaches to safeguard key systems and combat the dissemination of false information in the digital age.

References

Baltic Times. (2018) Lithuania's cyber security: key initiatives and future challenges. URL https://www.baltictimes.com/lithuania_s_cyber_security__key_initiatives_and_future_challenges/, accessed: 16/01/23.

Bankauskaite, D and Šlekys,D (2023) Lithuania's Total Defense Review, PRISM Vol. 10, No. 2, URL: https://ndupress.ndu.edu/Media/News/News-Article-View/Article/3323902/lithuanias-total-defense-review, accessed: 16/01/23.

BBC (2023) Has Putin's war failed and what does Russia want? https://www.bbc.com/news/world-europe-56720589, accessed: 16/01/23.

Bowen, A (2021) Russian Cyber Units – US Congressional Research Service, URL: https://crsreports.congress.gov/product/pdf/IF/IF11718, accessed 16/10/22.

Czosseck, C., Ottis, R., & Talihärm, A. (2011). Estonia after the 2007 Cyber Attacks: Legal, Strategic and Organisational Changes in Cyber Security. *International Journal of Cyber Warfare and Terrorism (IJCWT), 1*(1), 24-34. URL: http://doi.org/10.4018/ijcwt.2011010103

Debunk (NDa) About Debunk. URL: https://www.debunkeu.org/about, accessed: 16/01/23.

Debunk (NDb) Elves. URL: https://www.debunk.org/about-elves, accessed: 16/01/23.

International Telecommunication Union (2021). Global Cyber Security Index 2020, URL: https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2021-PDF-E.pdf. accessed: 16/01/23.

LRT (Lithuanian National Radio and Television) Russian propaganda targets Lithuania – Kaliningrad 'blockade' and 'betrayal, URL: https://www.lrt.lt/en/news-in-english/19/1736701/lrt-facts-russian-propaganda-targets-lithuania-kaliningrad-blockade-and-betrayal, accessed: 12/08/21.

Microsoft (2022a) Microsoft Digital Defence Report 2022, Microsoft.

Microsoft (2022b) Defending Ukraine: Early Lessons from the Cyber War, Microsoft.

Microsoft (2022c) Special Report: Ukraine – An overview of Russia's cyber attack activity in Ukraine, Microsoft.

Ministry of National Defence of the Republic of Lithuania (2017). National Cyber Security Strategy for 2017-2020. URL: https://kam.lt/download/87423/national_cyber_security_strategy_for_2017-2020.pdf, accessed: 12/08/21.

Ministry of National Defence of the Republic of Lithuania (2019). National Cyber Security Strategy.URL: https://kam.lt/wp-content/uploads/2022/11/2019-EN-KibernetineSaugumoStrategija-el.pdf, accessed: 16/01/23.

RCDC (ND) About RCDC (Regional Cyber Defence Centre), URL: https://www.nksc.lt/rkgc/en.html, accessed 2/2/23.

RCDC (2022a) RCDC 2022 1st Quarter Report, URL: https://www.nksc.lt/doc/rkgc/CTAC_2022_1st_Quarter_Report.pdf, accessed 2/2/23.

RCDC (2022b) RCDC 2022 2nd Quarter Report, URL: https://www.nksc.lt/doc/rkgc/CTAC_2022_2nd_Quarter_Report.pdf, accessed 2/2/23.

RCDC (2022c) RCDC 2022 3rd Quarter Report, URL: https://www.nksc.lt/doc/rkgc/CTAC_2022_3rd_Quarter_Report.pdf, accessed 2/2/23.

StopFake (2022) Fake: Lithuania Dismantles Railway to Kaliningrad, URL: https://www.stopfake.org/en/fake-lithuania-dismantles-railway-to-kaliningrad/, accessed 2/2/23.

Thales (2023) 2022-2023: A year of Cyber Conflict in Ukraine, Thales.

Time (2022) Meet the Lithuanian 'Elves' Fighting Russian Disinformation, URL: https://time.com/6155060/lithuania-russia-fighting-disinformation-ukraine, accessed 2/2/23.

US Government (2023) Pro-Russian Hacktivist Group 'KillNet' Threat to HPH Sector, Office of Information Security, URL: https://www.hhs.gov/sites/default/files/KillNet-analyst-note.pdf, accessed: 16/01/23.

Warren, M. (2020). Fake News Case Study during the Australian 2019 General Election. Australasian Journal of Information Systems, 24. https://doi.org/10.3127/ajis.v24i0.2803, accessed 2/2/23.

Warren, M, Štitilis, D and Laurinaitis, M (2023) The impact of Russian Cyber Attackers within the Ukraine Situation, Journal of Information Warfare, Volume 22, Issue 1.

Legal Response to Social Media Disinformation on National Level

MM Watney
University of Johannesburg, South Africa

mwatney@uj.ac.za

Abstract: Social media has an enormous impact on the manner in which society communicates and shares information. Digital is no longer a supplementary channel, but is the first place most people go to for news, information and communication. The transmission of social media disinformation has increased dramatically across the world and it necessitates a response. The discussion focuses on the response to social media disinformation on a national level. The discussion does not focus on foreign state or state-sponsored actors of misinformation. The focus and publicity may - within the context of cybersecurity - predominantly have been on cyberattacks, such as ransomware attacks. However, recent incidents - unrelated to foreign state interference and cyberattacks - illustrate that cybersecurity law must encompass the threat of disinformation. The 2020 COVID-19 pandemic, 2021 Washington, DC, United States, and South African as well as the 2023 Brazil riots illustrate the harmfulness of social media disinformation. Cognisance should be taken of the lessons learnt from the examples of social media disinformation as it may assist in determining a response to disinformation. There are various responses to national social media disinformation, such as legislative social media platform regulation, censorship, and criminalisation of the disinformation by itself. The response within the context of a cybersecurity threat landscape necessitates scrutiny as the response may impact on human rights. The trade-off between security and human right protection may be the violation of human rights to prevent harm from disinformation.

Keywords: Disinformation; misinformation; disinformation as a cybersecurity threat; social media platform regulation of disinformation; response to social media disinformation on national level.

1. Introduction

Facebook sees approximately 300 million new photos uploaded daily whereas six thousand tweets are published every second. The most popular YouTube channels receive over 14 billions views weekly, whilst the messaging app Telegram boasts over 500 million users (Hook and Verdeja, 2022). These statistics illustrate the enormous and unmistakable impact social media has on the manner in which society communicates and shares information and how deeply woven social media is into the fabric of society.

Social media has many societal benefits, such as providing a voice to many who have been voiceless in the past. However, social media communication and information-sharing can also be abused and exploited, as seen with the creation and distribution of disinformation. The concept of disinformation refers to false, inaccurate, or misleading information designed, presented, and promoted intentionally to cause public harm or make a profit (Colomina, Margalef, and Youngs, 2021).

Anyone with a social media account can create and spread disinformation, including governments, companies, other interest groups, or individuals. There are different actors responsible for disinformation (Colomina, Margalef, and Youngs, 2021), namely

- those who fabricate the disinformation, referred to as the instigators; and
- those who spread the content, referred to as the agents ('influencers', individuals, officials, groups, companies, and institutions).

Disinformation is aimed at driving a specific agenda. The motivation behind disinformation can be (Watney, 2018; Colomina, Margalef, and Youngs, 2021):

- personal (damage to a person or reputation of a company);
- political (to influence the public's viewpoint or ideology); or
- financial (to deceive).

Hook and Vereja (2022) indicates that researchers believe that organised social media disinformation campaigns have operated in at least 81 countries, a trend that continues to grow yearly.

Although the discussion does not focus on foreign interference with elections, Lasiello (2021) makes a valid point when he indicates that in 2016 the US government focussed predominantly on the threat of cyberattacks and "may have missed seeing the forest for the trees". Not only did the 2016 election reveal how ill-prepared the country was to address the soft-power tenets of information warfare (IW), but it also highlighted that social

media disinformation should form part of the cybersecurity threat landscape. Since 2016 there has been many incidents of social media disinformation unrelated to foreign interference and the attention has shifted to the cybersecurity threat that social media disinformation presents and the response thereto.

The aim of this discussion is to consider the manner in which parties, such as governments, social media users, private companies and service providers, may respond to social media disinformation. The main question will be discussed with reference to the following interlinked issues:

- Key terms will be conceptualised from a legal perspective such as misinformation, disinformation, fake news, propaganda, and cybersecurity.
- Reference will be made to examples of disinformation, such as the 2020 COVID-19 pandemic, the 6 January 2021 Washington, DC riot and the July 2021 South African riots.
- Disinformation and the impact of it will be highlighted within a cybersecurity threat landscape.
- The importance of a human rights centred approach to the mitigation of social media disinformation will be emphasised.
- The response to disinformation as a cybersecurity threat on a national level will be scrutinised.

2. Conceptualising terminology

It is important that terminology relevant to the discussion is conceptualised as it will serve as a point of reference with regards to the manner in which cybersecurity law may deal with social media disinformation nationally from a legal perspective. It appears that there are not universal definitions for the key concepts which is problematic as it may result in legal uncertainty.

The discussion focusses on the distribution of inaccurate or misleading communication and information by means of social media. “Social media” refers to computer-based platforms that enable people to communicate and share information across virtual networks in real-time. This includes community apps such as Facebook and Twitter, media-sharing channels like YouTube and Instagram, messaging apps such as WhatsApp, and hybrid platforms like Telegram (Hook and Verdeja, 2022).

A distinction is drawn between social media misinformation (SMM) — involving false or misleading information shared on these platforms — and social media disinformation (SMD), the subset of intentionally shared misinformation. The difference between misinformation and disinformation is that the latter is intentionally shared to cause harm. Malinformation on the other hand is true information which is shared to cause harm (Colomina, Margalef, and Youngs, 2021).

Understanding the difference between propaganda and disinformation is important as these terms are sometimes used interchangeably. Propaganda is information that might be true, but is often exaggerated, false, or misleading. It is used to influence public opinion to help a person, organization, or government. Propaganda creation is often associated with governments, but individuals and organizations outside of politics may use propaganda as well. The purpose of propaganda is to encourage people to promote a political cause or to think in a particular way and turn people away from a political cause (Bentzen, 2022).

Hook and Verdeja (2022) use the term *“social media misinformation*” and refer to SMD when there is a plausible indication of an intent to deceive such as with influence operations. In the discussion, the term “disinformation” is used as the discussion focuses on misleading or inaccurate information intentionally created and distributed on social media within a country and how the parties affected by such a cybersecurity threat should respond.

After the 2016 United States (US) elections, the term “fake news” was used, but experts have called for this term not to be used as it is too vague and ambiguous to capture the essence of disinformation (Colomina, Margalef, and Youngs, 2021). The term ‘fake news’ is commonly used to discredit the media. Since 2016 a clearer understanding of disinformation, has developed and for the purpose of legal certainty, the use of a more precise term such as disinformation is preferred. Baptista and Gradim (2022) propose the following working definition for “fake news”: “(1) a type of online disinformation, with (2) misleading and/or false statements that may or may not be associated with real events, (3) intentionally created to mislead and/or manipulate a public (4) specific or imagined, (5) through the appearance of a news format with an opportunistic structure (title, image, content) to attract the reader’s attention, in order to obtain more clicks and shares and, therefore, greater advertising revenue and/or ideological gain”. The latter definition will be applicable to the use of “fake news” in the discussion.

Disinformation should be seen within the context of cybersecurity. Cybersecurity is generally defined as the measures, policies, tools, and laws aimed at protecting hardware, software, and data in cyberspace against threats. Disinformation should be seen as part of the cybersecurity threat and risk landscape.

3. Examples of disinformation

The examples hereafter show that disinformation undermines human rights and many elements of good democratic practice (Colomina, Margalef and Youngs; 2021). Disinformation threatens freedom of thought, the right to privacy, and the right to democratic participation, as well as endangering a range of economic, social, and cultural rights. It also diminishes broader indicators of democratic quality, unsettling citizens' faith in democratic institutions by encouraging digital violence and suppression. On a global level, countries should take note of the lessons learnt from the social media disinformation examples in order to respond effectively to the threat and risk of disinformation.

Example 1: 2020 COVID-19 pandemic

Colomina, Margalef and Youngs (2021) indicates that during the pandemic, social media acted as a double-edged sword. On the one hand, digital platforms were useful in promoting debate within the scientific community and disseminating valuable information as well as investigative results. On the other hand, there was a great deal of disinformation. The disinformation was harmful as it resulted in mistrust in public institutions and put people's health at risk. The pandemic has also been referred to as an "infodemic" of misinformation. An infodemic refers to a large increase in the volume of information associated with a specific topic and whose growth can occur exponentially in a short period of time due to a specific incident. It should be noted that any big event may coincide with disinformation.

Example 2: Washington, DC riot

The 6 January 2021 Washington, DC, riot clearly illustrates the impact of disinformation during national elections (Almany *et al*, 2021; Jurecic, 2022). The riot followed a months-long disinformation campaign by former President Donald Trump and his allies, who claimed, without evidence, that the election had been stolen through fraud. The disinformation caused confusion and manipulation of social media users receptive to this rhetoric. The riot lasted seven hours, during which approximately 10,000 people came onto Capitol grounds, with many engaging in violent clashes with police officers trying to protect the building and lawmakers inside. At least 2,000 made it inside the Capitol building (Rubin, Mallin and Steakin, 2022).

Globally governments must be mindful of the impact of disinformation prior to and after an election. Increasingly it appears that elections on a national level may coincide with the use of social media disinformation aimed at causing political instability if a certain political party is not elected.

Example 3: Brazil

On 9 January 2023, Brazil experienced rioters storming democratic institutions, such as the Congress, Supreme Court, and Presidential Place. Brazilian analysts have long warned of the risk in Brazil of an incident similar to the 2021 insurrection at the US Capitol (Dwoskin, 2023). In the months leading up to the country's presidential election in October 2022, social media channels were flooded with disinformation, along with calls in Portuguese to "Stop the Steal" and cries for a military coup should the right wing incumbent, Bolsonaro, lose the election.

Example 4: 2021 South African riots

In July 2021, the province KwaZulu-Natal in South Africa (SA) ,experienced one of the worst uprisings since 1994. In 2022, **Media Monitoring Africa (MMA) director, William Bird told the SA Human Rights Commission that the disinformation posts shared during the July unrest sowed anger and distrust and contributed to the widespread violence and looting (**Pijoos, 2022)**.** While some social media messages served a genuine purpose to alert people to dangerous situations and places or roads to avoid, and areas to evacuate, many messages were misleading with out-of-context images or videos repurposed purely to be alarmist and deepen the sense of panic and anxiety (Malinga, 2021).

4. Challenges in mitigating social media disinformation

The characteristics of social media exacerbate the distribution and spread of disinformation (Watney, 2018). It provides for:

- instant and interactive communication;
- easy and cheap access;
- quick dissemination of disinformation to many; and
- a sense of interconnectedness.

Social media disinformation presents various inter-linked challenges, such as:

1. The rapid pace of social media and the technology culture associated with its rise accelerate the spread of misinformation, pushing it faster and further (Hook and Verdeja, 2022). The famous quote "A lie can travel around the world and back again while the truth is still lacing up its boots" resonates more than ever (Lapping; 2021). While professional journalists are trying to verify the facts before filing their reports, a deliberate falsehood created to stoke panic or anger can spread rapidly across social media. During COVID the term infodemic was used as this neologism (information + epidemic = infodemic) refers precisely to the speed with which disinformation spreads, similar to a virus (Baptista and Gradim, 2022).
2. Many actors of misinformation believe what they are sharing and are not intentionally spreading false information (Hook and Verdeja, 2022). There is a consensus that misinformation happens unintentionally (Colomina, Margalef and Youngs; 2021).
3. Not all harmful social media posts are wholly false. They may contain some truth, be largely true but misleading or out of context, or may advocate claims that are problematic but do not necessarily qualify as imminently dangerous.
4. Since social media provides the platform on which disinformation is spread, it needs to take responsibility for the content they distribute. The form of responsibility will be discussed hereafter at paragraph 5. It should be noted that the threat of disinformation must be mitigated by means of pro-active measures. Social media will always struggle to verify the truthfulness of content or assess the potential harm of each of the hundreds of thousands of posts their users generate every minute of the day. Social media platforms and their users must therefore share the responsibility of combating harmful or false content (Lapping, 2021).
5. There are many social media disinformation posts daily. It is not possible for law enforcement to respond to all social media disinformation postings. Determining the circumstances in which law enforcement must respond to social media disinformation as a cybersecurity threat may be problematic but the following lessons have been learnt from the 2021 Washington, DC, South African, and 2023 Brazil riots:

- Social media can be used as a tool to incite violence in the "real world". There are people who are looking to exploit and manipulate a situation by spreading disinformation and inciting violence to forward their own agendas (Malinga, 2021).
- The police should take a warning of impending violence seriously. For example, local officials, FBI informants, social media companies, former national security officials, researchers, lawmakers, and tipsters alerted law enforcement of possible violence on 6 January 2021, but law enforcement officials in Washington, DC did not respond with urgency to the cascade of warnings about possible violence (Almany *et al,* 2022). Washington, DC law enforcement may not have envisaged that social media communication could result in violence.

The 2021 US Capitol Hill riot following social media disinformation has now set a precedent and governments globally should be aware that elections on a national level may coincide with disinformation and possible protests and riots. Brazil law enforcement disregarded the warnings of possible violence.

6. Proving the origin of specific pieces of disinformation may be difficult (Hook and Verdeja, 2022).
7. It is difficult to track or completely prevent the spreading of disinformation. Content sharing has moved from open to encrypted platforms, such as WhatsApp.
8. There are many tools available to spread disinformation and it is not easy to keep up with all the technological tools used to impart disinformation. For example, the use of deep-fake technology can be harmful since people cannot tell whether content is genuine or false. The Cyberspace Administration

of China has started implementing China's new rules that prohibit the use of AI-generated content for spreading fake news or information deemed disruptive to the economy or national security (Hutton, 2023). In 2019 the US passed legislation to regulate AI-generative information concerning deepfakes and the EU is in the process of considering additional regulations for deepfakes and AI-generated imagery (Hutton, 2023).

5. Response to the cybersecurity threat of disinformation

5.1 Introduction

The challenges discussed at paragraph 4 shows that disinformation is a threat and requires pro-active measures.

Colomina, Margalef and Youngs (2021) identify three groups of actors in responding to disinformation, namely

- legislative and regulatory bodies;
- private sector such as social media platforms; and
- civil society.

Certain responses target the actors deemed responsible for disinformation; some target the disruptive techniques used whereas others aim at improving citizens resilience to disinformation.

The response must be human rights centred (Colomina, Margalef and Youngs, 2021). Although disinformation threatens human rights, the inverse challenge is that counter-disinformation policies can also restrict freedoms and rights. There may have to be a trade-off between security and the violation of the rights to freedom of expression and access to information but it will depend on the seriousness of the harm that the disinformation constitutes.

5.2 Threat actors held responsible for disinformation per se by means of anti-disinformation law

A complex issue is whether a government should implement anti-disinformation laws to hold the actors of disinformation criminally responsible. Those who support enacting anti-disinformation law criminalising disinformation argue that the right to freedom of speech does not include the freedom to spread disinformation. On the other hand, anti-disinformation law may violate freedom of speech and amount to censorship.

Different global contexts have diverse norms, ethical standards, and values around the blend of civil liberties, national security, and civilian protection as reflected by some countries that have criminalised disinformation *per se* (Hook and Verdeja, 2022).

In 2016, China criminalised the spread of information that may "undermine economic and social order". To further censor any possible misinformation, a law enacted in 2017 also demands that social media platforms must only circulate news from registered news outlets. Additionally, an app has been launched in China which allows its users to report those who are potentially spreading fake news (Elton, 2020).

Singapore enacted the Protection from Online Falsehoods and Manipulation Act (POFMA) in 2019. The POFMA strives to combat false statements of fact through online communication and provide support for the consequences of such communications. The POFMA also aims to censor false information that is against public interest. Contravening of the act may lead to a possible conviction of 10 year imprisonment and/or a $100,000 fine (Elton, 2020).

In 2022, Turkey adopted anti-disinformation law. Article 29 of the law provides that those who spread false information online about Turkey's security to "create fear and disturb public order" will face a prison sentence of one to three years. Sentences may be increased if the disinformation is spread through anonymous accounts (Shan, 2022).

In 2022, Russia lawmakers implemented a law imposing imprisonment of up to 15 years for spreading intentionally "fake" news about the military. They also impose fines for public calls for sanctions against Russia (Bloomberg News, 2022).

I opine that only the most serious disinformation content, such as the calls for incitement or the threat of violence to persons or damage to property based on the disinformation, should result in criminal liability. Criminalisation is a reactive measure and the emphasis should be on pro-active measures.

5.3 Social media platforms

5.3.1 Introduction

Social media platforms must take responsibility for disinformation as their platform is used for the dissemination of it, but the manner in which the responsibility is determined, is evolving (see par. 5.3.2)

There is a risk that the activities of the platforms in combating disinformation may restrict freedom of expression (Colomina, Margalef and Youngs, 2021). In 2021, Twitter banned former US president Trump from its platform amid concerns about Trump and his followers using social media to spread disinformation (Addler and Frier 2022; Jurecic, 2022; Lapping, 2021; Romano, 2021). Deplatforming former US president, Trump raised questions about limitations to freedom of expression and the role of gatekeepers of disinformation (Colomina, Margalef and Youngs, 2021). Romano (2021) indicates that deplatforming is effective at removing extremists from mainstream Internet spaces and he opines that it is not a violation of free speech, since free speech does not mean that there are no consequences or boundaries.

During the 2021 South African riots, a proposal was made to declare a state of emergency and to prevent access to certain communication towers in areas where the community was being mobilised to riot and loot (Nxumalo, 2022). However, social media can be used to access authentic information and serve as an efficient communication tool for urgent, ongoing and rapid up-to-date communication. By denying access to social media, the right to free speech and access to information would be violated unless the trade-off is security. There is a fine balance between security and protection of rights.

5.3.2 Evolution of social media self-regulation to legislative platform regulation

Platforms have terms and conditions that regulate the use of their platforms. Over the years attention has increasingly focussed on social media moderation and the responsibilities that platforms can and should have to control what people do on their sites.

Social media platforms have been attempting to address disinformation (Adler and Frier, 2022; Jurecic, 2022 Hetler, 2022). For example, Facebook runs two initiatives, News Integrity Initiative and Facebook Journalism Project, to highlight problems with disinformation and spread awareness. The platform also takes action against pages and individuals that share fake news and remove them from the site. Instagram and Facebook have a new "false information" label to combat disinformation. Third-party fact checkers review and identify potential false claims and posts. If this team determines this information is untrue, they flag it with a label to notify social media users it contains disinformation. When readers want to view a post with this label, they must click an acknowledgement that says the information is not true. If they try to share this information, they get a warning they are about to share disinformation (Hetler, 2022).

Many feel that social media companies are not doing enough in addressing disinformation. There is now a move away from platform self-regulation to platform legislative regulation to establish platform legal accountability pertaining to specifically illegal content. This move means that a social media platform can be held legally accountable if they do not comply with the rules.

Some governments have moved quicker to legal regulation than other governments. India put Twitter, Facebook, and the like under direct government oversight, enacting regulations requiring Internet platforms to help law enforcement identify those who post “mischievous information” (Adler and Frier, 2022). In China, Russia, and other countries subjected to authoritarian rule, governments actively censor the Internet, including blocking or greatly restricting access to American-owned social media sites (Adler and Frier, 2022).

The following legislation serves as examples of platform regulation:

- EU platform regulation in terms of the Digital Services Act (DSA)

On 5 July 2022, the European Union Parliament approved the DSA (https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en). The DSA gives member states new power to take down illegal content such as hate speech and terrorist propaganda. The platforms now have to comply with rules for content moderation by applying the so-called “notice and action mechanism”. If users contest a decision by the platforms regarding the illegality of user-generated content, they will have the possibility to seek judicial redress. In addition to illegal online content being removed, where it is criminal, it should be followed up by law enforcement and the judiciary. Online platforms will be obliged to report serious crimes to the competent authority. The DSA would help to address

harmful content (which might not be illegal) and the spread of disinformation by including provisions on mandatory risk assessments, risk mitigation measures, independent audits, and the transparency of so-called "recommender systems" (algorithms that determine what users see).

- United Kingdom (UK) platform regulation in terms of the Online Safety Bill (OSB)

Platforms will be required to (see https://www.gov.uk/government/publications/online-safety-bill-supporting-documents/online-safety-bill-factsheet):

- remove illegal content,
- remove material that violates their terms and conditions, and
- give users controls to help them avoid seeing certain types of content to be specified by the bill.

The OSB defines which conduct constitutes as illegal conduct and provides a general monitoring requirement for social media platforms. The main aim of the OSB is to remove illegal content and to protect children online. The social media platforms will continue mitigating against disinformation which does not fall under the concept of "illegal".

- United States reform of section 230 of the Communications Decency Act of 1996

Section 230 provides that an "interactive computer service" cannot be treated as the publisher or speaker of third-party content. This protects websites from lawsuits if a user posts illegal information, although there are exceptions for copyright violations, sex work-related material, and violations of federal criminal law. The spread of misinformation and disinformation through Internet services and denials of access to some individuals and materials have raised questions about the scope of Section 230 protections and many want reform of section 230 (Kozinets and Pheiffer, 2023; Hetler, 2022; Jurecic, 2022).

On October 3, 2022, the Supreme Court granted the petition for *Gonzalez v. Google*, an appeal from a decision of the United States Court of Appeals for the Ninth Circuit. The justices will weigh cases that could test legal protections for social media companies concerning posts on their platforms (Adler and Frier, 2022). The first case involves Google's alleged responsibility for terrorist propaganda on its subsidiary, YouTube. The plaintiffs argue that YouTube allegedly recommended ISIS content to users and therefore should not be protected by section 230 for its own algorithmic recommendation.

5.4 Private companies or organisations

Companies and organisations should anticipate that they may be victims of disinformation campaigns and show people how to verify content about their business or organisation (Lapping 2021). For example, a disinformation campaign specifically targeting the Black Lives Matter (BLM) organisation resulted in BLM calling on the public to assist with disinformation against the movement (see https://blacklivesmatter.com/help-us-fight-disinformation/). The disinformation may be a consequence of the 2020 BLM protests that erupted globally after the death of George Floyd on May 25, 2020 under the knee of a Minneapolis police officer, Derek Chauvin. The protests were also characterised by a great deal of mis- and disinformation which caused confusion, stoked tension, and distracted from factually correct information (**Georgacopoulos and Poche, 2020).**

5.5 Cybersecurity resilience

Cybersecurity is the ability to withstand, respond to, and recover from all types of information and communication technology (ICT) related disruptions and threats which include the threat of disinformation. A government has to identify the threat and risk that disinformation presents and determine how it will respond. For example, the manner in which the South African government and law enforcement responded to the 2021 riots was inadequate. There should have been up-to-date communication on a consistent basis during the riots.

6. Conclusion

There is a fine line between balancing social media's potential use for good and bad. Unfortunately, the characteristics of social media assist with the creation and spread of disinformation.

Social media disinformation forms part of the cybersecurity threat and risk landscape on national level. The take-away from the discussion is that disinformation will not abate and is exacerbated by certain events. Disinformation requires pro-active measures to mitigate against it.

Governments, social media platforms, organisations, and social media users are in the early stages of responding to disinformation. They should consider the possible outcomes of their response carefully. The response to disinformation can easily result in the violation of free speech contrary to the essence of social media, such as freedom of expression and access to information.

References

Adler, M. and Frier, S. (2022), "Twitter, Musk and Why Online Speech Gets Moderated", [online], https://www.washingtonpost.com/business/twitter-musk-and-why-online-speech-gets-moderated/2022/10/03/0cb0ae68-434f-11ed-be17-89cbe6b8c0a5_story.html.

Alemany, J. *et al*; (2022) "Jan. 6 insurrection: Washington Post investigation of the cost, causes and aftermath" [online], https://www.washingtonpost.com/politics/interactive/2021/jan-6-insurrection-capitol/.

Baptista, J.P. and Gradim, A. (2022) "A working definition for fake news", [online], https://www.mdpi.com/2673-8392/2/1/43.

Bentzen, N. (2022) "Understanding propaganda and disinformation", [online], https://epthinktank.eu/2015/11/17/understanding-propaganda-and-disinformation/.

Bloomberg News, (2022) " Russia criminalises sanctions calls and fake news on military", [online], https://www.bloomberg.com/news/articles/2022-03-04/russia-to-punish-sanctions-appeals-and-fake-news-on-military?leadSource=uverify%20wall.

Colomina, CL, Margalef, HS. and Youngs, R. (2021) "The impact of disinformation on democratic processes and human rights in the world", [online], https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf.

Dwoskin, E. (2023) "Come to the 'war cry party': How social media helped drive mayhem in Brazil" **[online],** https://www.washingtonpost.com/technology/2023/01/08/brazil-bolsanaro-twitter-facebook/.

Elton, W. (2020) "Is it illegal to post fake news?", [online], https://zegal.com/en-nz/blog/post/is-it-illegal-to-post-false-information/.

Georgacopoulos, C. and Poche, T. (2020) "Fake news, disinformation and the George Floyd protests", [online], https://faculty.lsu.edu/fakenews/about/protestfakenews.php.

Hetler, A. (2022) "10 ways to spot disinformation on social media", [online], https://www.techtarget.com/whatis/feature/10-ways-to-spot-disinformation-on-social-media.

Hook, K. and Verdeja, E. (2022) "Social Media Misinformation and the Prevention of Political Instability and Mass Atrocities", [online], https://www.stimson.org/2022/social-media-misinformation-and-the-prevention-of-political-instability-and-mass-atrocities/.

Hutton, C. (2023) "China implements new rules to regulate 'deepfakes' and AI art", [online], https://www.washingtonexaminer.com/policy/technology/china-implements-new-rules-over-deepfakes-chatgpt-ai-art.

Jurecic, Q. (2022) "The politics of Section 230 reform: Learning from FOSTA's mistakes", [online], https://www.brookings.edu/research/the-politics-of-section-230-reform-learning-from-fostas-mistakes/.

Kozinets, R. and Pheiffer, J.(2023) "Beyond Section 230: A Pair of Social Media Experts Describes How to Bring Transparency and Accountability to the Industry", [online] https://www.nextgov.com/ideas/2023/01/beyond-section-230-pair-social-media-experts-describes-how-bring-transparency-and-accountability-industry/381532/.

Lapping, G. (2021) "The July riots: an inflection point for digital media", [online], https://www.businesslive.co.za/redzone/news-insights/2021-09-01-the-july-riots-an-inflection-point-for-digital-media/.

Lasiello, E. (2021) "What is the Role of Cyber Operations in Information Warfare?", [online], https://digitalcommons.usf.edu/cgi/viewcontent.cgi?article=1931&context=jss.

Malinga, S. (2021) "Social media used as a dangerous tool to mobilise ongoing attack", [online], https://www.itweb.co.za/content/DZQ58MVPpK3vzXy2.

Nxumalo, L. (2022) "July unrest: social media fuelled unrest" [online], https://www.iol.co.za/sunday-tribune/news/july-unrest-social-media-fuelled-unrest-481bbafe-1ce1-4ca7-87c9-7da6cfc652eb.

Pijoos, I. (2022) "Harmful disinformation sowed anger, violence, and anarchy during July unrest - media expert ", [online], https://www.news24.com/news24/southafrica/news/harmful-disinformation-sowed-anger-violence-and-anarchy-during-july-unrest-media-expert-20220304.

Romano, A. (2021) "Kicking people off social media isn't about free speech", [online], https://www.vox.com/culture/22230847/deplatforming-free-speech-controversy-trump.

Rubin, O, Mallin, A. and Steakin, W. (2022) "By the numbers: How the Jan. 6 investigation is shaping up 1 year later", [online], https://abcnews.go.com/US/numbers-jan-investigation-shaping-year/story?id=82057743.

Shan, L.Y. (2022) "Turkey has long been hell for journalists': Reporters slam country's new 'fake news' law", [online], https://www.cnbc.com/2022/10/21/turkish-reporters-slam-countrys-new-fake-news-law.html.

Watney, M.M. (2018) "The Legal Position of Social Media Intermediaries in Addressing Fake News" **European Conference on Cyber Warfare and Security; Reading**, [online], https://books.google.co.za/books?id=-kFmDwAAQBAJ&pg=PA497&lpg=PA497&dq=murdoch+watney+fake+news&source=bl&ots=ZERXYgrf_n&sig=ACfU3U3XdbTGA27HmOT6PLnpmg_Om7W4HA&hl=en&sa=X&ved=2ahUKEwiExJWR_q38AhXOesAKHSERA68Q6AF6BAgkEAM#v=onepage&q=murdoch%20watney%20fake%20news&f=false.

On Benchmarking and Validation in Wargames

Adam J. Wilden[1], Mehwish Nasim[1,2,4,*], Peter Williams[3], Tim Legrand[4], Benjamin Peter Turnbull[5] and Patricia A. H. Williams[1]

[1]Flinders University, Adelaide, Australia
[2]The University of Western Australia, Perth, Australia
[3]Defence Science and Technology Group, Adelaide, Australia
[4]The University of Adelaide, Adelaide, Australia
[5]University of New South Wales, Australian Defence Force Academy, Canberra, Australia

adam.wilden@flinders.edu.au
mehwish.nasim@uwa.edu.au
peter.williams2@defence.gov.au
tim.legrand@adelaide.edu.au
b.turnbull@adfa.edu.au
trish.williams@flinders.edu.au

Abstract: There are multiple arguments for and against wargames. Many scientists do not recognise the science in wargames. It is suggested that there is not enough literature relating to wargaming, for there to be any large-scale research into wargames. This is primarily because scientists often refuse to publish results, thus creating a vicious cycle where research is not published because there is not enough research being published. This ultimately deters researchers from studying wargaming in any serious fashion. Owing to this limitation, published work on the results, and protocols of wargames are scarce in scholarly research. Wargaming has considerably less academic focus with a fragmented and practical focus on design and benchmarking. This is surprising given the long history of wargaming (dating back to the early 1600's), when compared to the relatively recent history of other domains such as software engineering. To better understand the current state of research into wargaming in reference to benchmarking and validation, a scoping review (SR) was conducted. The scholarly research into wargaming reveals papers on general modelling, conflict modelling, influence modelling, evaluation of wargames, analytical tools, use of AI in wargame design, evaluation of predictive modelling in wargames, improving command and control through wargaming, and cost-benefit analysis for decision making. The initial analysis of the coverage of wargaming research, together with the limited number of papers found, indicate that there is a distinct lack of academic research into wargaming. Additionally, there is a wide variety of areas that are interested in the wargaming field, however, with no universal method of analysis or benchmarking, this limits the reproducibility of results, and the ability to judge the overall effectiveness of wargaming efforts. Wargame designers need to be able to assess wargame components to validate, compare, and predict the effects on gameplay and for decision-makers to draw conclusions with more confidence.

Keywords: Validation; Benchmarking; Reproducibility; Wargaming; Information Operations

1. Introduction

Wargames have long been in history. They were used to communicate ideas, understand trends, predict future outcomes, create strategies, test models and plans, and for training. The success or failure of a wargame lies in its ability to enable participants to express possible reactions and its connection to storytelling. The Military Operations Research Society (MORS) defines a wargame as a model or simulation involving two or more opposing forces using data, rules and procedures designed to depict an actual or fake narrative. Other organisations such as NATO define wargame as the use of a data processing system to represent selected behavioural characteristics of a physical or abstract system.

The earliest uses of first-generation (board game style) wargames are estimated at around 2300BC, while the use of second-generation (military simulation style) wargames is estimated at around the early 1600's (Caffrey 2019). Wargames can serve multiple purposes, including communication of ideas, understanding trends, predicting future outcomes, creating strategies, testing models and plans, and so on. The success or failure of a wargame lies in its ability to enable participants to express possible reactions and its connection to storytelling.

Compared to the long history of wargaming, research into other areas such as software engineering (which is comparatively a relatively recent phenomenon), has a vast amount of available published research. Therefore, it is surprising that there is not the same level of research into standardisation and efficiency of wargames as there is for software engineering.

1.1 Wargame Research

When considering investigating the efficacy of utilising research into wargaming for academic or serious professional use, it is worth considering the arguments for and against wargames, such as those mentioned in Wojtowicz (2019). It is suggested that many scientists do not recognise the science in wargames due to the lack of sufficient literature to support research on wargames. However, this can become a vicious cycle, if scientists refuse to publish wargaming results, due to wargaming results not being available.

The wargaming publication paradox often deters researchers from pursuing wargaming as a serious research discipline. This makes finding published scholarly work on results, protocols of wargames as well as dedicated researchers in the field hard to find for scholarly research.

2. Scoping Review on Benchmarking and Validation in Wargaming

We conducted a review on the scholarly literature available on benchmarking and validation in wargames.

2.1 Methodology

A Scoping Review (SR) was selected as it provides a reliable result in a reasonable amount of time when compared to a full Systematic Literature Review (SLR). This is supported by research by Munn et al. (2018), where it is suggested that a scoping review can suit a broader research focus and allow a better inclusion criterion for subject areas in which the exact nature of the subject is not well published in an academic sense, which is the case for wargaming.

In addition to taking considerably longer to perform, an SLR would be more appropriate for situations where there is a large number of published academic work, with which to be able to explain the current best practice (Fink 2019). The SR methodology published by Arksey & O'Malley (2005), with additional information and improvements to the methodology from Levac et al. (2010), was the chosen methodology for this research.

2.1.1 Research Question

This SR focuses on providing an overview of the current state of literature related to wargaming adjudication and benchmarking. Therefore, the following three Research Questions (RQ), were formulated to identify who benefits from further research into wargaming adjudication (RQ1), what is involved in adjudication and benchmarking (RQ2), and what sort of challenges are presented in the literature (RQ3). These are provided in Table 1.

Additionally, researchers will be able to use these findings to help better identify limitations and gaps in the literature, and potentially untapped future research in the wargaming domain.

Table 1: Research Questions

Number	Question	Motivation
RQ1	Who are the stakeholders that would benefit from a more standardised form of wargaming adjudication and overall benchmarking?	To identify the types of people/groups that will benefit from the increase in research into benchmarking and adjudication.
RQ2	What are the types of tools that are used in wargaming adjudication benchmarking	To help the researchers identify what sort of tools are used within the wargaming domain regarding adjudication and benchmarking.
RQ3	What are the reported challenges in employing adjudication benchmarking and/or standardisation that need to be addressed in further research?	To identify the current state of the wargaming domain; what is the current level of research? what challenges are mentioned (if any)? what are the gaps in research? and are there any future directions mentioned?

2.1.2 Search Strategy

The search process was conducted by two authors and was done independently using the SCOPUS abstract and citation database. For results that were unavailable in SCOPUS in full-text form, the GOOGLE Scholar web search engine and full-text database was used to source full-text versions of the documents for review.

The search string used to retrieve the documents was designed to utilise title, keyword, and abstract, to identify candidate papers that are related to Wargaming. The search process took place between February 2022 and April 2022. The review process was performed by the same authors individually. Abstracts were reviewed to identify relevant articles, and those that were not relevant to the subject were discarded. Additional information

was gathered from the SCOPUS search results, such as document type, document tags (key words), publication date, publication source, funding affiliation, and subject area.

2.1.3 SCOPUS Search String

The SCOPUS advanced search allowed the use of Boolean operators and regular expressions, in addition to limiting the search to specific areas of the document, such as title, abstract, and key words. The search was limited to papers published in 2012 or after to reduce the initial number of results.

The wargame search component was searched using the ”TITLE-ABS-KEY” function, to limit the search to only the title, abstract, and keywords sections. This was decided as being necessary due to the frequent use of the term wargame in the body of many unrelated papers. Whereas the use of wargame in the title would suggest a focus on wargaming as the subject, and not in passing. The use of the wildcard operator ’*’ allows zero or more unspecified characters to appear after the section of quoted text. This allows us to target both the singular use of wargame, and the activity wargaming, in one search term.

”War gam” OR ”Wargam*”*

The second component of both searches was the inclusion of capability, maturity, and adjudication into the search terms. This was enhanced with the use of the wild card operator in the same fashion as the previous search term, to broaden the use of all of the required words.

”Capabilit” OR ”Maturity” OR ”Adjudicat*”*

The final component was that of analysis and benchmarking into the search terms. Again, this was done in the same fashion as shown above.

”Assess” OR ”Analy*” OR ”Benchmark” OR ”Bench mark” OR ”compar*” OR ”Study” OR*

”Investigat”*

2.1.4 Study Selection

The two primary authors applied the SCOPUS advanced search strings mentioned in Section 2.1.3, to select the final set of papers used in the Scoping Review, the search results retrieved from SCOPUS are shown in Table 2. Further refinement of the selection was based on the criteria in Table 3.

Table 2: Full SCOPUS Search Results

ID	Authors	Title	Year
PID1	[No author name available]	Bridging multiple autonomous vehicle disciplines: Ensuring autonomous systems promote peace and stability on the world's oceans	2016
PID2	Athmer K., Gaughan C., McDonnell J.S., Leach R., Davis B., Truong K., Borum H., Leslie R., Ma L.	Enhancing army analysis capability for warfighter protection: TRADOC-RDECOM M&S decision support environment collaboration	2012
PID3	Banerjee P.	The wisdom of the road: Research and pedagogy on India-China and the Silk Roads Ethos (SRE)	2018
PID4	Barrett R., Deger-Sen S., Sen S.	Rational model of conflict: War and peace in South Asia	2013
PID5	Chen L., Kou Y., Li Z., Xu A., Wu C.	Empirical research on complex networks modeling of combat SoS based on data from real war-game, Part I: Statistical characteristics	2018
PID6	Christensen K.D., Dobias P.	Wargaming the use of intermediate force capabilities in the gray zone	2021
PID7	Colbert E., Sullivan D., Kott A.	Cyber wargaming on SCADA systems	2017
PID8	Colbert E.J.M., Kott A., Knachel L.P.	The game-theoretic model and experimental investigation of cyber wargaming	2020
PID9	D'Antonio C., Gower S., Young A., Teague E.	Non-kinetic operations for stabilizing government	2014
PID10	Davis P.K., Bracken P.	Artificial intelligence for wargaming and modeling	2022
PID11	De Lima Filho G.M., Kuroswiski A.R., Medeiros F.L.L., Voskuijl M., Monsuur H., Passaro A.	Optimization of Unmanned Air Vehicle Tactical Formation in War Games	2022
PID12	DeBerry W.T., Dill R., Hopkinson K., Hodson D.D., Grimaila M.	The wargame commodity course of action automated analysis method	2021
PID13	Dewoody R., Hayes J., Isnardi N., Kalinka J., Teague L.T.C.E.	Irregular warfare models for regional stability development	2014
PID14	Duong D., Bladon C.	Interfacing and validating models of the US army TRAC tactical war game	2012

ID	Authors	Title	Year
PID15	Duong D.V.	The design of computer simulation experiments of complex adaptive social systems for risk based analysis of intervention strategies	2012
PID16	Eichberg H.	Questioning play: What play can tell us about social life	2016
PID17	Fridheim H., Malerud S.	"Shall we play a game?" the pros and cons of wargaming reachback support in security risk analysis	2020
PID18	Fridheim H., Malerud S.	“Shall We Play a Game?” The Pros and Cons of Wargaming Reachback Support in Security Risk Analysis	2020
PID19	Gambone M.D.	Small wars: Low-intensity threats and the American response since Vietnam	2012
PID20	Goecks V.G., Waytowich N., Asher D.E., Jun Park S., Mittrick M., Richardson J., Vindiola M., Logie A., Dennison M., Trout T., Narayanan P., Kott A.	On games and simulators as a platform for development of artificial intelligence for command and control	2022
PID21	Grogan P.T., De Weck O.L.	Infrastructure System Simulation Interoperability Using the High-Level Architecture	2018
PID22	Guarda T., Vaca O.B., Pinguave M.P., Maldonado E.P., Augusto M.F., Orozco W., Pinto F.M.	Wargames applied to naval decision-making process	2017
PID23	Guo R.J., Sprague K.	Replication of human operators’ situation assessment and decision making for simulated area reconnaissance in wargames	2016
PID24	Hart B., Hart D., Gayle R., Oppel F., Xavier P., Whetzel J.	Dante agent architecture for force-on-force wargame simulation and training	2017
PID25	Hernandez A.	Advancements in scenario-Based learning through quantification of wargaming techniques	2017
PID26	Hernandez A.S.	Integrating simulation-driven decisions and business wargames to shape fiscal policies	2015
PID27	Hixson J.G., Miller B., May C.	Enhanced modeling and simulation of EO/IR sensor systems	2015
PID28	Hocaoğlu M.F.	Agent-based target evaluation and fire doctrine: an aspect-oriented programming view	2022
PID29	Hocaoğlu M.F.	Rule based target evaluation and fire doctrine	2019
PID30	Hodický J., Procházka D., Baxa F., Melichar J., Krejčík M., Křížek P., Stodola P., Drozd J.	Computer assisted wargame for military capability-based planning	2020
PID31	Jensen R., Teo G., Lunsford J., DeFalco J.A.	Designing a Distributed Team Training Experience to Balance Data Collection Needs With Natural Team Interaction	2021
PID32	Johnson B.	Predictive analytics in the naval maritime domain	2020
PID33	Kaivo-Oja J.	Three theoretical approaches to pirate entrepreneurship: Towards future studies of pirate entrepreneurship	2014
PID34	Kanniainen V.	Cyber Technology and the Arms Race	2019
PID35	Keller J., Colosi C., Nichols G.	Achieving high performance aircraft modeling in operational analysis software using a minimum complexity model	2018
PID36	Kim S.-J., Tsuruoka T., Hasegawa T., Aono M., Terabe K., Aono M.	Decision maker based on atomic switches	2016
PID37	Kodalle T., Sample C., Ormrod D., Scott K.	Thoughts about a general theory of influence in a dime/pmesii/ascop/irc2 model	2019
PID38	Largent M., Jensen G., Law R.	The Design for Maritime Singularity: Exploration of Human/AI Teaming and Organizational Carrying Capacity for the U.S. Navy	2018
PID39	Leonard W.B., Rinaudo C.H., Richards J.E., Goerger S.R.	An analytical model for system-of-systems sensor detection and defense efforts	2020
PID40	Mittal V., Davidson A.	Combining Wargaming with Modeling and Simulation to Project Future Military Technology Requirements	2021
PID41	Najgebauer A.	War games, simulation and CI tools for strategic planning	2020
PID42	Ormrod D., Scott K., Scheinman L., Kodalle T., Sample C., Turnbull B.	The persuasion game: Developing a serious game based model for information warfare and influence studies	2019
PID43	Rao D., Hu X., Wu L.	Performance analysis of parallel data distribution management in large-scale battlefield simulation	2014
PID44	Rossetti M.D., Bright J.	Bulk petroleum supply chain simulation modeling	2019
PID45	Rowe C., Seif Zadeh H., Garanovich I.L., Jiang L., Bilusich D., Nunes-Vaz R., Ween A.	Prioritizing investment in military cyber capability using risk analysis	2019
PID46	Ruvinsky A., Gonzalez M.E., Richards J., Salter C., Church J., McKee C.D., Seale M.	An Approach to Gamifying Acquisitions for Assessing Impact on Military Strategy of Nation States	2021

ID	Authors	Title	Year
PID47	Schwartz P.J., O'Neill D.V., Bentz M.E., Brown A., Doyle B.S., Liepa O.C., Lawrence R., Hull R.D.	AI-enabled wargaming in the military decision making process	2020
PID48	Schwarz J.O., Ram C., Rohrbeck R.	Combining scenario planning and business wargaming to better anticipate future competitive dynamics	2019
PID49	Sun P., Zhang J., Wang L.-H.	The Application Study on Accurately Search&Rescue of the Wounded on the land battlefield base on "beidou + Armored ambulance	2021
PID50	Vienhage P., Barcomb H., Marshall K., Black W.A., Coons A., Tran H.T., Nguyen T.M., Guillen A.T., Yoh J., Kizer J., Rogers B.A.	War-gaming application for future space systems acquisition: MATLAB implementation of war-gaming acquisition models and simulation results	2017
PID51	Wei X., Yang L., Cao G., Lu T., Wang B.	Recurrent MADDPG for object detection and assignment in combat tasks	2020
PID52	West J., Chu M., Crooks L., Bradley-Ho M.	Strategy war games: how business can outperform the competition	2018
PID53	Yu A., Yick K.L., Ng S.P., Yip J.	Case study on the effects of fit and material of sports gloves on hand performance	2019

Table 3: Inclusion (I) and Exclusion (E) criteria

Inclusion or Exclusion	*Criteria*
E1	Topic too narrow, or on a niche subject.
E2	Does not have a full-text available
I1	Must be related to Benchmarking and/or Adjudication in Wargaming

The SCOPUS search returned 53 documents published between 2012 and 2022. Of those 53, in the initial selection pass, 15 were discarded, with 13 discarded based on the I1 criteria, and 2 being discarded due to being a duplicate entry of an existing search result. These papers (PIDs 1, 2, 3, 15, 16, 17, 21, 27, 33, 36, 41, 47, 48, 50, 53) are shown in Table 2.

During the second pass, where there was greater attention placed on the content, and relation to the Scoping Review Topic. During this review process, 23 additional papers were discarded leaving 14 papers.

Of the 23 additional papers discarded, six were discarded based on E1 for being too narrow in focus (PID 5, 22, 25, 31, 37, 51), and 17 papers were discarded due to failing I1 (Topic too specific), after further investigation into the content of the paper.

The topics that these papers were primarily focused on were the following: Geo-Political Conflict (PID4), UAVs (PID11), Remotely Run Wargames (PID18), COA Predictive Analysis (PID32), Military Acquisitions (PID46), Ambulances (PID49), Target Evaluation (PID29), AI/Technology Singularity (PID38), Economics and Business Wargames (PID 26, 34, 44, 45, 48, 52), technology Roadmaps (PID40), Intrusion Detection (PID39), Data Distribution Analysis (PID43). These documents are listed in Table 2.

2.1.5 Data extraction

Data extraction was performed by one author, with a pre-defined data extraction form (**Error! Reference source not found.**) to record the extracted data gathered from the selected papers. The SCOPUS advanced search feature gave access to all the required data (such as document meta-data, search tags, and full-text files). Additionally, and Analyse Search Results options available in SCOPUS, such as Research Affiliation, Publication Location, and Study Sponsors.

Table 4: General Data Extraction Form

Item	Question	Related RQ
Stakeholders	What sponsors, research affiliates, government agencies, and militaries are related to the paper?	RQ1
Domain	What is the publishing domain that the paper is related to?	RQ1
Tags	What are the search tags of the paper?	RQ1
Tools	What are the wargaming tools/software that are mentioned?	RQ2
Challenges	What are the challenges mentioned (if any)?	RQ3
Future	What are the remaining research questions mentioned in the paper (if any)?	RQ3

3. Preliminary Scoping Review Result

3.1 Demographics

After filtering out unrelated documents, our remaining dataset contained 14 papers published between 2012 and 2022, with the largest number of yearly publications occurring in 2021 and 2022. We can see that there is a steady increase in publications related to Wargame Adjudication and Benchmarking (Figure 1). This increase in publications over time could suggest that there is an increasing interest in wargaming, especially from 2020 onward. However, due to the lack of data to go off (with the small number of overall publications), it is hard to say for certain that this is indeed an increase in interest.
Of the documents retained during the SR process, the document type was mostly Journal Articles (8), with slightly less Conference papers (5), and one Book Chapter.

3.2 Tools

The common tools (Table 5) that were discussed in many papers appeared to be primarily free tools (RStudio), or tools that are commonly available (such as Microsoft Excel, or Google Docs/Slides). Another common factor was the use of AI, modelling, and simulation tools, to evaluate decision making.

Table 5: Data Extraction Form for Wargaming Tools

Tool(s)	Related Papers
Hybrid Wargame Environment using Webex, Google Docs, and Google Slides.	(Christensen & Dobias 2021)
SCADA (Supervisory Control and Data Acquisition) systems.	(Colbert, E, Sullivan & Kott 2017)
Game-theoretical Framework.	(Colbert, EJM, Kott & Knachel 2020)
Regression Modelling in RStudio.	(D'Antonio et al. 2014)
Machine Learning AI, and Agent Based Modelling.	(Davis & Bracken 2022)
Wargaming Commodity Course of Action Automated Analysis Method (WCCAAM).	(DeBerry et al. 2021)
Microsoft Excel, Crystal Ball, Vensim, Simulation Models (Lanchester Model, Monte Carlo Simulation).	(Dewoody et al. 2014)
Impact Computing's open-source model integration framework XBM, BayesOwl probabilistic inference engine.	(Duong & Bladon 2012)
OpSim, Deep Learning Tools.	(Goecks et al. 2022)
Dynamic Replicator of Decision Making (DRDM) Tool.	(Guo & Sprague 2016)
Dante (built on the Umbra Simulation Framework).	(Hart et al. 2017)
AdSiF (Agent-driven Simulation Framework).	(Hocaoğlu 2022)
Microsoft Excel.	(Hodickỳ et al. 2020)
Custom Applications (simulated news feed, social media, common operating picture), Slack, NoSQL databases.	(Ormrod et al. 2019)

3.3 Future Work

Table 6 shows the extracted summary of the reviewed papers future work sections. It is worth mentioning that some papers did not include any reference to future work (PID20, and PID28).

Table 6: Extraction Form for Future Work

Future Work	Description	Related Papers
Additional Research	Research into different scenarios. Adjusting initial assumptions, to test different strategies.	(Christensen & Dobias 2021), (Colbert, EJM, Kott & Knachel 2020)
Additional Complexity	Adding in additional complexity to the model. Extend to whole planning process, allow more input from users.	(Colbert, E, Sullivan & Kott 2017), (Hodickỳ et al. 2020)
Application of Model	Incorporating Model into a Wargame.	(D'Antonio et al. 2014)
Further use of AI	Further implementation of AI/Machine learning, particularly Cognitive Modelling. Replacing various processes with machine Learning. Implementation of automated online experimentation.	(Davis & Bracken 2022), (DeBerry et al. 2021), (Hart et al. 2017)
More Data	Research into operational data to replace Model data. Expanded collection of data, development of bots.	(Dewoody et al. 2014), (Ormrod et al. 2019)
Refining Processes	Fine tuning the processes to better measure performance and enhance consistency.	(Guo & Sprague 2016)

Table 7 shows that the primary publication source is the United States of America (9 publications), then Canada (2 publications), with the rest of the locations having one publication each. It is not surprising that the primary

publication location is the United States of America, given their large budget for military programs. The rest also coincide with large military budgets and needs.

Table 7: Scoping Review Document Source

Country / Territory	Documents
United States	9
Canada	2
Australia	1
Czech Republic	1
Germany	1
Turkey	1
United Kingdom	1
Undefined	1

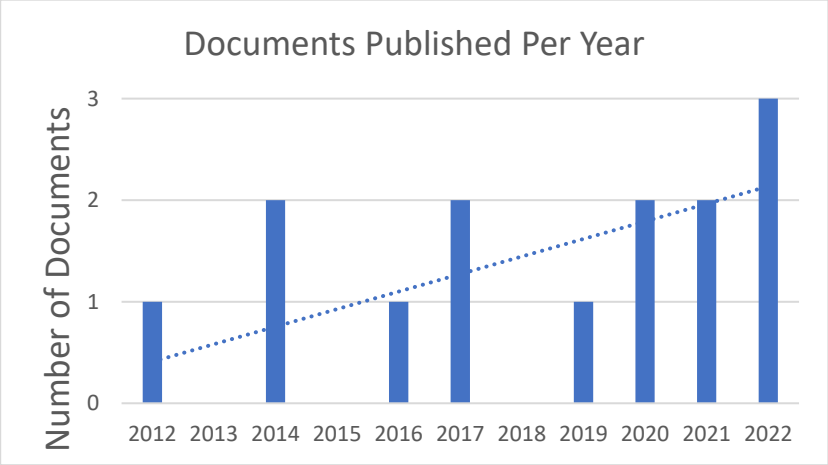


Figure 1: Scoping Review Documents Published Per Year

As we can see from Figure 1, There is a relatively low number of publications, until around 2020. Where the Journal of Defense Modelling and Simulation increased related publications from one per year, to three in 2022 (Figure 2). Suggesting that in addition to the overall increase in papers related to wargaming adjudication and benchmarking, the output per publisher increased over that period.

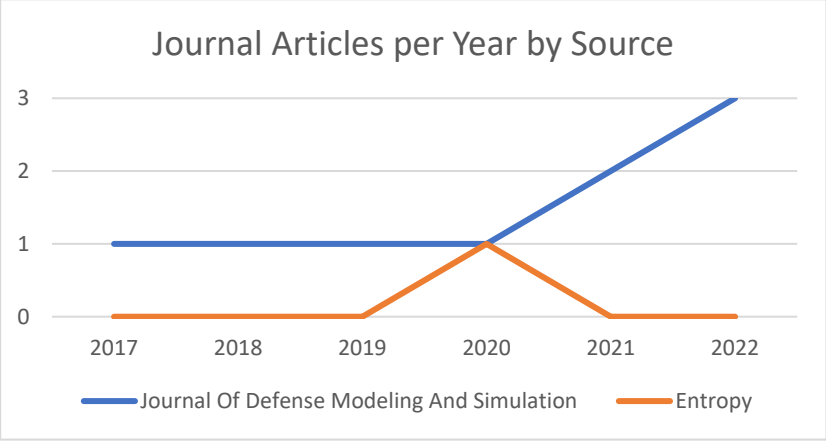


Figure 2: Scoping Review Journal Articles per Year by Source

As we can see from Table 8, the primary stakeholders who would benefit the most are world militaries (primarily the US, with 8 documents being associated with various US Army/Air Force departments). Private research laboratories, such as Defence Research and Development Canada, CAE, and RAND are the next largest group of stakeholders. Additionally, there were a selection of global universities, such as Yale, De Montfort, UNSW

Sydney, and Istanbul Medeniyet. Finally, military universities such as Command and Staff College of the German Armed Forces Univerzita obrany v Brne, and United States Military Academy at West Point.

Table 8: Scoping Review Document Research Affiliation

Affiliation	Description	Gov/Mil/Private	Documents
U.S. Army Research Laboratory	Foundational research laboratory.	Military (US)	5
Defence Research and Development Canada	Advisor, collaborative partner, and knowledge integrator for Canadian defence and security.	Private	2
NATO Headquarters Supreme Allied Commander Transformation	One of NATO's strategic commanders and commanding officer of Allied Command Transformation.	Military (NATO)	1
Agent Based Learning Systems	Global high technology company.	Private	1
CAE Defence & Security	Global high technology company.	Private	1
Communications-Electronics Research Development and Engineering Center	In-house science and engineering	Military (US)	1
Command and Staff College of the German Armed Forces	Military College.	Military (German)	1
Univerzita obrany v Brne	Military College.	Military (Czechia)	1
RAND Corporation		Private	1
Yale University	University	Private	1
Sandia National Laboratories, New Mexico	Global high technology company.	Private	1
De Montfort University	University	Private	1
Air Force Institute of Technology	Air Force academic education research, and continuing education.	Military (US)	1
United States Military Academy at West Point	US Army Officer education.	Military (US)	1
UNSW Sydney	University	Private	1
Istanbul Medeniyet University	University	Private	1

3.4 Domains

The remaining papers retrieved in the Scoping Review suggests that there is a wide variety of subject areas in which wargaming is being published. However, as can be seen in Table 9, most of the published work is found in the Engineering, Mathematic, and Computer Science domains, and the remaining research domains contribute a relatively small number of publications each.

Table 9: Scoping Review Research Disciplines

Subject area	Documents
Engineering	12
Mathematics	7
Computer Science	5
Arts and Humanities	1
Decision Sciences	1
Physics and Astronomy	1

3.5 Document Tags

The search keywords used for the publications of the SR showed a wide variety of different research areas, as shown in the word cloud in Figure 3. Of the 84 tags used in the documents. The ones that received the most uses were Modelling (7), Math (4), AI (3), and Simulation (3).

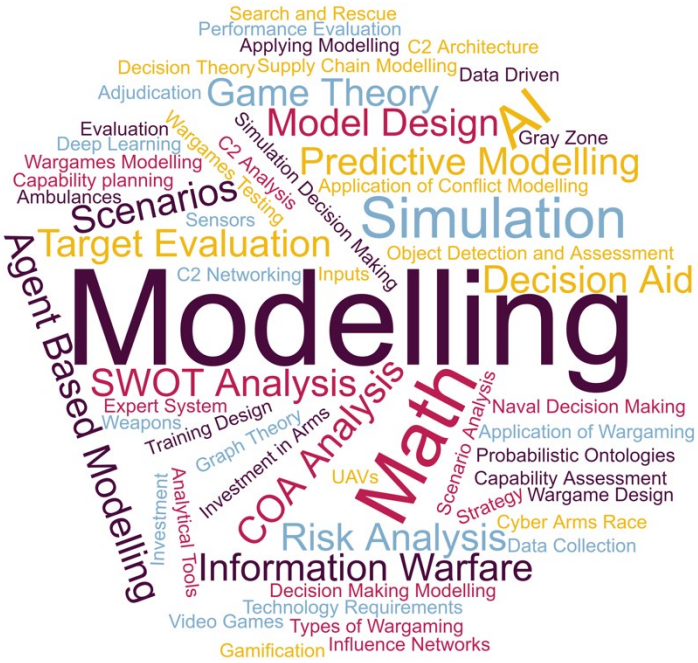


Figure 3: Document Tag Word Cloud

3.6 RQ1 Who are the stakeholders that would benefit from a more standardised form of war gaming adjudication and overall benchmarking?

The research shows that the US military may be the most interested in wargaming research. With most publications being facilitated by one of the many branches, and research departments of the US Armed Forces. Private research facilities and universities would also be stakeholders, with many private research corporations providing research support. There appears to be good coverage of public and private organisations, however, there most definitely is a bias towards the United States with regards to research.

3.7 RQ2 What are the types of tools that are used in war gaming adjudication benchmarking?

The thematic analysis on the tools used in wargaming revealed 4 categories:

1. Simulation and modelling: These tools are used to create virtual representations of battlefield scenarios and to predict the outcomes of different courses of action. The tools include Hybrid Wargame Environments, SCADA systems, Game-theoretical Frameworks, Regression Modelling, Machine Learning AI, Agent-based Modelling, Wargaming Commodity Course of Action Automated Analysis Methods, Vensim, Simulation Models, OpSim, Dante, AdSiF, XBM, BayesOwl, Custom Applications.
2. Decision Making: These tools are used to model the decision-making processes of the different actors involved in a wargame scenario. They can be used to analyse the relative strengths and weaknesses of different strategies and tactics. The tools include Game-theoretical Frameworks, Regression Modelling, Wargaming Commodity Course of Action Automated Analysis Methods, Dynamic Replicator of Decision-Making Tools.
3. Data Analysis and Management: These tools are used to analyse large amounts of data generated by wargame simulations and to present it in a clear and understandable format. These are software systems that provide decision-making support to the wargame adjudicator by integrating data, models, and analytical tools. They help the adjudicator evaluate the different options and to make informed decisions. These tools include SCADA systems, Regression Modelling, Machine Learning AI, Wargaming Commodity Course of Action Automated Analysis Methods, Microsoft Excel, Crystal Ball, OpSim, Deep Learning Tools, XBM, BayesOwl, Custom Applications, Slack, and NoSQL databases.
4. Collaboration and Communication: These tools are used to facilitate communication and collaboration among the wargame participants. These tools include Hybrid Wargame Environment using Webex, Google Docs, and Google Slides or Custom Applications (simulated news feed, social media, common operating picture), Slack, NoSQL databases.

3.8 RQ3 What are the reported challenges in employing adjudication benchmarking and/or standardisation that need to be addressed in further research?

Of the documents that remained in the SR, the future directions involved such subjects as Additional Research (Christensen & Dobias 2021), Adding additional complexity to the computer system being researched (Colbert, E, Sullivan & Kott 2017; Hodickỳ et al. 2020), application of non-kinetic operation models to test suitability (D'Antonio et al. 2014), the extended incorporation of AI and Machine Learning for better optimisation of processes (Davis & Bracken 2022; DeBerry et al. 2021; Hart et al. 2017), collecting data to better model real systems and incorporate a further collection of data in systems (Dewoody et al. 2014; Ormrod et al. 2019), and refining overall processes to improve performance (Guo & Sprague 2016).

What is in common between most of the future directions mentioned, was the necessity for the acquisition of more data, and more research into the area. Research into the systems and processes mentioned, together with the use of tools, will require a significant amount of data to either train or validate results.

4. Discussion and Future Work

From the analysis of coverage of wargaming research, and the small number of papers related to both wargaming, and adjudication/benchmarking (described in section 2.1.4), it can be shown that there is a distinct lack of total academic research into the area. There is a trend towards more research, especially from 2020 onwards, but it is still a relatively small number of publications.

The results show that the domains of publication are not restricted to one area, and there is a substantial institutional interest in developing wargames. However, even with the increased interest, and available research funding from a variety of governmental sources (primarily in the US), there still appears to be no universal method of analysis or benchmarking.

Further research will need to be performed into three areas, first what is the progress of publications into the field after the initial SR. Second, to expand the scope of research by identifying what types of wargames specifically are the most documented with regards to adjudication and benchmarking. Finally, continue assessment of the type of wargames that would benefit the most from a standardised form of adjudication and benchmarking, and what stakeholders prefer these types.

Acknowledgment

*This research was supported by the South Australian Defence Innovation Partnership – Collaborative Research Funds, awarded to Dr Mehwish Nasim.

References

Arksey, H & O'Malley, L 2005, 'Scoping studies: towards a methodological framework', *International journal of social research methodology*, vol. 8, no. 1, pp. 19--32.

Caffrey, MB 2019, *On wargaming: How wargames have shaped history and how they may shape the future*, vol. 43.

Christensen, KD & Dobias, P 2021, 'Wargaming the use of intermediate force capabilities in the gray zone', *Journal of Defense Modeling and Simulation*.

Colbert, E, Sullivan, D & Kott, A 2017, 'Cyber wargaming on SCADA systems', paper presented to Proceedings of the 12th International Conference on Cyber Warfare and Security, ICCWS 2017, <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85018965308&partnerID=40&md5=2057ac3bb92b99a6db1c8d1de873d87c>.

Colbert, EJM, Kott, A & Knachel, LP 2020, 'The game-theoretic model and experimental investigation of cyber wargaming', *Journal of Defense Modeling and Simulation*, vol. 17, no. 1, pp. 21-38.

D'Antonio, C, Gower, S, Young, A & Teague, E 2014, 'Non-kinetic operations for stabilizing government', paper presented to 2014 IEEE Systems and Information Engineering Design Symposium, SIEDS 2014, <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84903710483&doi=10.1109%2fSIEDS.2014.6829899&partnerID=40&md5=9f071bea57b5804809396d058f984c83>.

Davis, PK & Bracken, P 2022, 'Artificial intelligence for wargaming and modeling', *Journal of Defense Modeling and Simulation*.

DeBerry, WT, Dill, R, Hopkinson, K, Hodson, DD & Grimaila, M 2021, 'The wargame commodity course of action automated analysis method', *Journal of Defense Modeling and Simulation*.

Dewoody, R, Hayes, J, Isnardi, N, Kalinka, J & Teague, LTCE 2014, 'Irregular warfare models for regional stability development', paper presented to 2014 IEEE Systems and Information Engineering Design Symposium, SIEDS 2014,

<https://www.scopus.com/inward/record.uri?eid=2-s2.0-84903718259&doi=10.1109%2fSIEDS.2014.6829898&partnerID=40&md5=2bb462b3f21bab6924e2ffe81c592160>.

Duong, D & Bladon, C 2012, *Interfacing and validating models of the US army TRAC tactical war game*.

Fink, A 2019, *Conducting research literature reviews: From the internet to paper*.

Goecks, VG, Waytowich, N, Asher, DE, Jun Park, S, Mittrick, M, Richardson, J, Vindiola, M, Logie, A, Dennison, M, Trout, T, Narayanan, P & Kott, A 2022, 'On games and simulators as a platform for development of artificial intelligence for command and control', *Journal of Defense Modeling and Simulation*.

Guo, RJ & Sprague, K 2016, 'Replication of human operators' situation assessment and decision making for simulated area reconnaissance in wargames', *Journal of Defense Modeling and Simulation*, vol. 13, no. 2, pp. 213-25.

Hart, B, Hart, D, Gayle, R, Oppel, F, Xavier, P & Whetzel, J 2017, 'Dante agent architecture for force-on-force wargame simulation and training', paper presented to Proceedings of the 13th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2017, <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067103347&partnerID=40&md5=8b3b6657a74dd965a2bd85d8e96b3641>.

Hocaoğlu, MF 2022, 'Agent-based target evaluation and fire doctrine: an aspect-oriented programming view', *Journal of Defense Modeling and Simulation*, vol. 19, no. 1, pp. 107-21.

Hodickỳ, J, Procházka, D, Baxa, F, Melichar, J, Krejčík, M, Křížek, P, Stodola, P & Drozd, J 2020, 'Computer assisted wargame for military capability-based planning', *Entropy*, vol. 22, no. 8.

Levac, D, Colquhoun, H & O'Brien, KK 2010, 'Scoping studies: advancing the methodology', *Implementation science*, vol. 5, no. 1, pp. 1--9.

Ormrod, D, Scott, K, Scheinman, L, Kodalle, T, Sample, C & Turnbull, B 2019, 'The persuasion game: Developing a serious game based model for information warfare and influence studies', paper presented to European Conference on Information Warfare and Security, ECCWS, <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85070013506&partnerID=40&md5=871656af04f1918e9452ab4f4c8aee6d>.

Wojtowicz, N 2019, 'From sandboxes to laboratories: evolution of wargaming into a method for experimental studies', *International Journal of Scientific and Research Publications*, vol. 9, no. 12, p. 399.

The UN Global Digital Compact (GDC), Achieving a trusted, free, open, and Secure Internet: Trust-building

Allison Wylde
Glasgow Caledonian University (London)

allison.wylde@gcu.ac.uk

Abstract: A United Nations' (UN) public consultation, underway, is reviewing requirements for the Global Digital Compact (GDC) to advance UN goals for an 'open, free, and secure digital future for all' (UN, GDC, 2022). Achieving the goals relies on proposed principles, including: connecting everyone; avoiding fragmentation; protecting data; applying human rights; accountability for discrimination and misleading content; regulation of artificial intelligence; digital commons as a public good; and 'other' areas. The purpose of this paper is to present an argument that trust must be included as a central 'other' principle. Although successful achievement of the GDC goals is contingent on building trust in each principle, a method for trust-building is not provided. Through leveraging well-established organization and conflict management trust-building literature the contribution of this paper presents a fresh conceptual framework, allowing trust and trust-building in the goals to be operationalized and assessed. In, addressing the research gap as to how build trust in the GDC goals as they are implemented, the novel trust-building process as presented helps policymakers, practitioners, and academics better address potential risks to the future internet, such as, increased; state isolation, sovereignty, and internet fragmentation. Limitations and calls for further research highlight that understanding state-level trust-building in policy is not yet mature. Further, scholars needs to better categorize the processes, dynamics and norms involved in state-level trust-building, helping to counter future internet challenges.

Keywords: United Nations (UN), Global Digital Compact (GDC), internet, trust, institutional trust, cyber security, digital futures, future internet

1. Introduction

The UN has highlighted "stark and urgent problems facing all of humanity" and the "breakdown of trust and solidarity in one another" (UN, CA, 2021, p.4). In response they propose a Global Digital Compact (GDC) to achieve an open, free, and secure internet (UN, GDC, 2022). Eight principles are proposed: connecting everyone; protecting data; applying human rights; digital commons as a public good; regulation of artificial intelligence; accountability for discrimination and misleading content; avoiding fragmentation; and 'other' areas (UN, GDC, 2022); which provide a list of dos and don'ts, and a theoretical basis (public goods) to achieve goals. This paper argues that as the GDC goals are founded on trust, and the UN calls for trust-building (UN RDC, 2020: UN, CA, 2021), trust should be incorporated as a central integrating principle.

This paper argues that by positioning trust as a central and integrating principle; the goals can be delineated and evaluated. The contributions of this paper, propose an argument that trust be included as the 'other' principle, and address the question of building trust; well-established organizational and conflict resolution trust-building theory are leveraged to allow evaluation of the implementation of the GCU goals.

This section, the introduction, is followed by section 2, which explores, key challenges in East-West states' interpretations of norms, sovereignty and internet fragmentation, as examples of internet weaponization; followed by a discussion on the UN's calls for trust, followed by possible scenarios for the future internet. Section 3, reviews trust-building theory; with a focus on how trust can be understood, leveraged, and operationalized to assess goals are implemented. Section 4 presents an analysis of trust. Finally, in section 5, the conclusion, contribution, limitations, and promising directions for future studies together with the paper's implications are presented.

2. Achieving an open, free, and secure internet: the challenges and the goals

Key challenges face the internet, arguably acting as drivers behind the establishment of the GDC goals. What follows is not a systematic state of the art, rather an attempt to reach a concise understanding with the focus on relevant material.

2.1 Challenges to our digital futures

Since the 1980's when the web emerged the ethos has shifted from open-source and benevolence to increased deliberate, malicious actions, including state-sponsored hacking, criminal behaviour, harassment and attacks on the internet's structure (The Guardian, 2019): resulting in weaponizing the internet (Henriksen, 2019).

Prominent examples include the actions of advanced persistent threat actors (APTs); WannaCry and Not Petya, and breaches such as SolarWinds.

2.1.1 Actions of state-actors

Challenging relations among major state-actors including China, Russia and the US/EU/UK have emerged, arguably related to state doctrine. From 2013, in high-level government experts' meetings (GGE), the UN has called for international laws, human rights and fundamental freedoms (freedom of expression) to apply in ICT (GGE, 2013: GGE, 2015); alongside promoting cyber norms for predictability (hence trust) in the internet and ICT (Henriksen, 2019). UN's proposed cybercrime convention, in development for 2024, is challenged by some states. Russia opposes including human rights in the convention, suggesting that while human rights are enshrined in fundamental documents of the UN, individual states' must have rights take measures in accordance with their domestic laws (UNDOC, 2023a). This approach appears consistent with China and Russian previous actions in rewriting existing normative frameworks (Henriksen, 2019); considered next.

Russia as a GGE member promotes ideas such as the primacy of states in maintaining secure and peaceful ICT environments, territorial integrity (UNDOC, 2023a), and national information space (Claessen, 2020). Russia also contests UN norms and is attempting to centralize control and sovereignty through state legislation (Russian State Duma, 2018, in Claessen, 2020). By amending existing laws (Epifanova, 2020), Russia seeks to limit, register, and secure internet traffic through the creation of a duplicate infrastructure (Claessen, 2019). Amendments include compulsory installation of threat countermeasures, border control for connection lines crossing the Russian border and a Russian register of domain names (Epifanova, 2020). Russian also uses surveillance and isolation of its internet, as precedents for other states (Epifanova, 2020: Claessen (2019). Use of the term information rather than cyber by Russia, enables information security to encompass information-psychological and information-technical: arguably, a cyber arms race, to seek arms-control agreements (Henriksen, 2019). Current challenges by these state-actors and their allies, to the UN's proposed 2024 cybersecurity convention (UN, 2023) could result in a dilution of the provision for human rights, confusion over definitions and weakened capacity for responses to criminal actions (GDP, 2023). At the same time, China has sought to normalize human rights violations through rewriting human rights normative frameworks from obligations (Henriksen, 2019) to ideas that human rights are negotiable (Currie, 2023).

For the EU, the distributed denial of services (DDoS) attacks on Estonia shifted the EU's view from global multi-stakeholder cooperation for internet oversight to a position of global cyber security (EU, 2011, in Claessen, 2019; Schmidt, 2013). The EU propose a common approach to critical infrastructure security and engagement through public-private partnerships (Claessen, 2019).

Arguably the different approaches by states have culminated in UN calls for goals of an open, free, and secure digital future for all (UN, GDC, 2022). Key content in the calls includes trust and trust-building , trust in digital identities and integrity in finance and tax, set out in Table 1, below, and discussed next.

2.2 UN responses: calls for trust

Responding to challenges facing the internet, the UN has proposed policy actions, measures, and principles, all of which cite trust. Given limitations in scope, this discussion clarifies the calls for trust from the UN Secretary-General's Roadmap for Digital Cooperation (UN, RDC, 2020); Our Common Agenda (UN, CA, 2021); and, finally, the Technical Envoy's Office, Global Digital Compact (UN, GDC, 2022). Although UN Sustainable Development Goals (SDGs) framework, notably, SDG 16, Peace, Justice, and Strong Institutions and, SDG 17, Partnerships for the Goals (UN, SDG, n.d.) and the forthcoming 2024 UN cybersecurity convention (UNOCD, 2023b) are acknowledged; this paper confined to examining UN calls for trust and trust-building (UN, RDC, 2020: CA,2021: UN, GDC, 2022).

Drawing on earlier commitments the 2015 GGE aims to maintain peace stability, promoting an open, secure, peaceful, and accessible ICT environment based on non-binding state behaviour, founded on norms (Assembly, UG, 2015; GA, 2013, cited in Broeders et al., 2021). The framework is includes (1) no-interference with the public core of the internet (2) protecting electoral services (3) avoiding tampering (4) non commandeering of ICT devices (botnets) and (5) reduction and mitigation of significant vulnerabilities (UN, SG Panel, 2019). As an observation, these norms are largely limiting norms (UN, GDC, 2022).

Policy on preventative measures was replaced by ideas of good practices and positive duties (UN, GDC, 2022). In 2020, during the UN's 75th anniversary, a Roadmap for Digital Cooperation was released; based on, connect,

respect, and protect, with actors adhering to goals of: an inclusive digital economy and society; global connectivity; digital public goods; and, digital inclusion (UN, RDC, 2020). These actions promote open-source software, open data, open AI models, open standards, open content and the adherence to laws, standards, and best practices (UN, RDC, 2020). Specialist groups established to support the goals, includes, the digital public goods alliance (digitalpublicgoods.net, 2022) to promote positive actions including monitoring connectivity, digital inclusion, capacity building, trust-building, and oversight of AI (UN, RDC, 2020).

Subsequently in 2021, the Common Agenda aimed to protect the online space and strengthen governance (UN, CA, 2021). The Common Agenda's Commitments include: (1) leave no one behind; (2) protect our planet; (3) promote peace and prevent conflict; (4) abide by international law and ensure justice; (5) place women and girls at the centre; (6) build trust; (7) improve digital cooperation; (8) upgrade the UN; (9) ensure sustainable financing; (10) boost partnerships; (11) listen to and work with youth; and, (12) be prepared (UN, CA,2021).

Specifically, in the CA, commitments aim to improving digital cooperation, ensure that governance keeps pace and protecting online space, and through formation of the Internet Governance Forum, to innovate, reform and support effective governance of the digital commons (UN, CA, 2021). The key trust related actions, goals and principles in the policies are summarized below to allow comparison across the three policies (Table 1, below).

In summarizing the comparison of trust statements in the three policies (UN, RDC, 2020: CA,2021: UN, GDC, 2022); the trust elements are largely in parallel. In contrast with the earlier limiting behaviours, calls for non-binding norms seek to build on good practices and focus on positive duties (Table, 1, above). Examples of trust aims include building global cooperation, commitment (notably, UN RDC, 2020) and working in partnership (CA,2021). However, the policy documents are silent on the practical steps to achieve these aims.

Although minor differences are acknowledged, though, building trust (UN RDC, 2020: CA,2021: UN GDC, 2022) and enhancing trust (UN, RDC, 2020) as compared with tackling and addressing trust (CA,2021: UN, GDC), the scope of three documents are consistent in terms of trust, integrity, digital identity technology and silence on guidance on the processes involved in operationalizing trust and trust-building (UN RDC, 2020: CA,2021: UN GDC, 2022).

2.3 Possible futures for the internet

As discussed earlier in section 2.2, continued current tensions between competing state doctrines result in challenges to the internet, including: fragmentation of the internet, human rights abuses, state surveillance and election interference. What may happen next in terms of the future of the internet can be viewed through three scenarios and resulting actions: increased polarization towards state sovereignty; increased internet fragmentation; potentially countered by, entrepreneurs promoting and facilitating trust-building in cyber norms such as the GDC. Given space constraints the next section is confined to the possible impacts of each scenario.

Should current trends towards state sovereignty persist, state-actors harden their laws and norms of political and military responses, risking spillover into 'above-threshold' actions that could prompt military responses.

Table 1. Comparison of the key trust statements (summarized) in the Roadmap for Digital Cooperation (DC, 2020), the Common Agenda (CA,2021) and the Global Digital Compact (UN, GDC, 2022).

UN, RDC (2020)	UN, CA (2021)	UN, GDC (2022)
n=15 [21]	n=43 [47]	n=4
Promote engagement with and prioritise digital trust	Build and rebuild trust	Build trust and address, digital trust, and insecurity
Build global commitment to and enhance trust, trust in technology, in digital identity and in trustworthy AI	*Build integrity, *Improve experience with public infrastructure and basic services, *Inclusive listening, *Tackle corruption, *Reform tax, *Financial integrity	Inequality leads to mistrust between countries and mistrust in institutions
Prioritize trust and security	work in partnership to build trust	address digital trust and security
Note, n = equals the number of times trust (in context) appears in each document; where a [number] is shown, this is the total observations, including material that has been excluded as not relevant to the context.		

Continued internet fragmentation can be viewed as impacting across two levels. At the level of inter-state relations, fragmentation could result in state isolation. At the level of citizen, state actions could interfere in election, or through surveillance of citizens, could drive down rights and freedoms. These impacts longer term could lead to further deteriorating relations.

Conversely, should the influence of norm entrepreneurs' increase, impacts of the former two scenarios could be reduced. As norms evolve increased levels of cooperation and trust results, in part due to familiarity and increased utility (Wylde, 2022). The benefits could reduce friction between states generating increasing returns on engaging in cooperation.

2.4 Summary

From the perspective of the UN GDC, the dominance of policy, goals and principles rests on positive actions, the "dos" as compared with the single "don't" principle, avoiding internet fragmentation. Secondly, and in contrast, current and future challenges highlight emerging state calls for sovereignty and moves towards internet fragmentation. Arguably and thirdly, this demonstrates the need for norm entrepreneurs to cooperate and strengthen a message of the economic utility of trust between states. Whether this future approach based on cooperation will develop trust (or not) remains unclear. The next sections reviews trust-building in the GDC goals (UN, GDC, 2022).

3. Achieving the GDC goals: trust-building theory

As highlighted in Table 1 above, although the selected policy documents all refer to trust, building trust and trustworthiness (as well as related terms), are not explicit. In the discussion that follows, prominent trust and trust-building theory is discussed, followed by the first steps towards a conceptual framework to allow an assessment of trust in policy to be undertaken.

3.1 Trust and trust-building theory

Although trust is well researched and conceptualized among organization and management scholars (notably Mayer et al., 195; Rousseau et al., 1998) and from conflict resolution (Deutsch, 1958), this paper argues that in policymaking, little guidance is provided on definitions of trust and the operationalization of trust-building. In addressing this research gap, understanding leveraging, and operationalizing trust theory is presented. Given space limitations, a state of the art is not presented, rather a focused discussion on key theory provides a framework for evaluation.

Trust is a construct that operates in and at different levels across different entities, for example, at the level of individual trustor-trustee relations, or trust in teams, or trust at an organizational or institutional level (Fulmer and Gelfand, 2012). Trust also exits beyond trustor' relations, as trust in technology (McKnight, 2011) and trust in institutions (Bachman and Inkpen, 2011: Rousseau et al., 1998). The paper aims to understand trust in institutional policies, principles, and trust in 'applying human rights' (UN, GDC, 2022). Institutional trust is based on favourable assumptions by a trustor of a trustees' future behaviour. Trustors' trust institutions, their norms, and patterns of behaviour (Bachmann and Inkpen, 2011). Institution-level trust, provides organizational supports (such as legal systems), reduces vulnerability for trustors (Rousseau et al., 1998) and relies on trusted people in organizations (Vanneste, 2016)

This paper defines trust as based on positive expectations by a trustor that a trustee will perform an action, irrespective of trustor' control or monitoring, trust requires a trustor to accept vulnerability (Mayer et al., 1995: Rousseau et al., 1998); moderated by a trustor's propensity to trust (Mayer et al., 1995), ability to trust and trust experience (Deutsch, 1958). This definition allows for evaluation of antecedents, processes, and outcomes of trust (Wylde, 2021); the conceptual framework is presented next.

3.2 Trust analysis: conceptual framework

Trust, in an institutional concept, such as a goal or a principle, is next evaluated, through drawing on two major approaches from trust-building theory, organization and conflict studies. The aim is to evaluate trust in UN goals, principles, positive actions and best practices along with one limiting principle (avoiding internet fragmentation) and theory (UN, GDC, 2022).

The frame of analysis draws from the well-established integrative trust model first proposed by Mayer at al. (1995), in organization and management studies together with Deutsch's (1958) foundational work on trust in conflict resolution studies. The strengths of these approaches lie in their ability to provide a process-based view to assessing the construct of trust-building, in sometimes, contested relations, essential given the complexities of managing complex public problems (Oomsells and Bouckaer 2014). Utility of this approach is shown in recent Google Scholar searchers, terms the "Integrative trust model" returned about 2.75m results, and "Conflict resolution AND trust" returned about 2.6m (GoogleScholar.com, 08 Jan. 2023). The analysis framework separates, trust-antecedents, trust-assessments, and the post-trust actions (Table 2, below). Firstly, trust antecedents include positive expectations of a trustee's trustworthiness, moderated by propensity to trust (Mayer at al., 1995), plus the ability to trust, moderated by prior trust experiences (Deutsch, 1958). The trust assessment, second, evaluates the trustee's ability, benevolence, and integrity (ABI) (Mayer et al., 1995). ABI is explained as ability (capability to undertake a specific task), benevolence (goodwill), and finally, integrity (honesty). Thirdly, following trust, a trustor's accepts vulnerability and takes a risk, as set out in table 2 below.

The strengths of this conceptual process framework lie in its ability to disentangle complex, dynamic and interlinked elements into a process allowing trust to be assessed. Further, by conceptualizing trust as occurring at an institutional level, vulnerability and risk are reduced (Mayer et al., 1995); through trusting policy (Bachman and Inkpen, 2011: Rousseau et al., 1998) and the organizations' people (Vanneste,2016).

Table 2. Trust-building process model, viewed from the perspective of a trustor (source the author, developed from Mayer et al., 1995: Deutsch, 1958: Wylde 2021)

Antecedents to trust	Trust assessment	Post-trust actions
Ability to trust	Ability	Acceptance of vulnerability
Trust, prior experience	Benevolence	
	Integrity (ABI)	
	Assessment of risk	Risk-taking behavior
Mediated by propensity	Mediated by propensity	Mediated by propensity

4. The goals and principles: assessing trust

In achieving a free, open and secure digital future eight principles are proposed for stakeholders (UN, Tech Envoy, 2022); the conceptual framework proposed in table 2 is applied.

Given the complexities in managing societal issues (Oomsells and Bouckaer 2014) this paper adopts a process-based approach to disentangle the separate elements of trust (Wylde, 2021) to help understand what is involved in securing the UN goals. What follows next is a simplified set-by-step process to the assessment of trust in the key trust elements as summarized in Table 1. This is followed by a trust assessment of the state of trust in three key state-actors involved in UN GGE talks, China, Russia, and the US.

4.1 Trust assessment

As a start, the antecedents of trust are provided by UN calls for an open and secure internet (UN, RDC, 2020 CA, 2021), reliant on trust, building trust and trustworthy frameworks (Table 1). These examples provide evidence for the presence of presumptive trust and an ability to trust in the goals. At the same time, stakeholders cooperate through partnerships to achieve these goals; demonstrating a propensity to trust.

In the next stage assessment of trust, the propensity to trust is assumed to reflect presumptive trust on the part of the trustor (Table 2). A trustor assesses the policy or principle based on its ability (can the policy achieve its task?) benevolence (do the policy intentions demonstrate goodwill towards stakeholders?) and integrity (does the policy demonstrate honesty) cf., Mayer et al. (1995). From conflict resolution considerations include prior experiences of trust and ability to trust (Deutsch, 1958); supported by institutional trust (Bachmann and Inkpen, 2011: Rousseau et al., 1998: Vanneste, 2016)

If the ABI elements are assessed favourably, the final step is trust formation, with a caveat, as trust requires a trustor's acceptance of vulnerability (Table 2). In institutional trust, trustors are predisposed to trust institutions (and their people), thereby reducing vulnerability (Bachmann and Inkpen, 2011: Rousseau et al., 1998: Vanneste, 2016); resulting in an increased likelihood of trust in the policy or principle.

4.2 Operationalizing the assessment of the trust-building

An assessment state of trust among three key players, China, Russia, and the US is presented next. In examining the antecedents of trust, state-actors Russia and America can be considered as sharing a common Cold War history and experience while China, experienced a cultural revolution. These experiences may act as an antecedent, predisposing these states to a propensity that is closer to distrust, trust and verify (Lewicki, 1998) and zero trust (Wylde, 2021).

In terms of the ability to trust, although the US, UK and other western democracies hold free elections, Russia, Iran, and North Korea are accused with election and referendum interference (O'Connor et al., 2020) resulting in a loss of trust in the election process (O'Connor et al., 2020). Actors experiencing election interference may have their ability to trust negatively impacted.

Fukuyama considers that culture is also important, for example, individuals in China, may be more likely to be predisposed to trusting family above state-actors, arguably due to experiences of distrust in state-actors and institutions (Fukuyama, 1996).

As institutional trust is in question; actors may be considered to possess favourable predispositions, though impacted negatively through prior experiences.

4.3 Summary

By simplifying the steps involved in assessing trust in a goal or a principle through drawing on the process model (Table 2). An assessment of trust among the major state-actors can be undertaken. Pointing to the value in leveraging well-established trust theory to disentangle the complex processes and dynamics involved in trust. Once trust is assessed, any deficit in trust can be identified and recommendations for trust-building can follow, based on the steps outlined. To the best of the author's knowledge, this is the first attempt at assessing trust in policies (goals and principles) as operationalized, evaluated and monitored.

5. Conclusion

The contribution and limitations of the paper, followed by promising directions for future studies, and the implications, are presented next.

In addressing questions of trust in policy, this paper examined and reviewed important challenges faced by the internet together with implementation of UN goals/ principles, and possible future internet scenarios. The paper presents an argument that to achieve the GDC goals, trust must be included as a central principle.

In examining key challenges to the internet: emerging state calls for sovereignty and moves towards internet fragmentation were identified. The paper highlighted the importance of cooperation between norm entrepreneurs in norm implementation along with communication of the economic utility of state-to-state trust.

The research gap was addressed through the contribution which leveraged well-established trust and conflict management theory to create a conceptual framework. This allowed the complex processes and dynamics involved in trust-building to be disentangled. Next, the framework was operationalized to evaluate trust in policy among the major state-actors. Findings suggest that actors may starting from a position of distrust (or zero trust). The paper argues trust assessment allows gaps in trust to be identified and recommendations for trust-building provided, therefore further demonstrating the importance in understanding the steps involved in trust-building and categorizing the processes involved.

These first steps towards a trust-building conceptual framework (Table 2) aim to provide a consistent approach to assessing and categorizing trust. Recognising the multifaceted processes and elements involved in trust and trust-building the contribution helps as policy is operationalized, and trust is evaluated and monitored.

As with all research limitations are inevitable, for example, given time/ space limitations, the evaluation focused on three linked UN policies. Nevertheless, the work presented here suggested that UN policy is inter-linked. Through consideration of potential scenarios for the future internet, important directions for future research suggest the importance of examining potential future state responses and actions. This would allow better understanding of the phenomena of trust-building among state-actors. Interesting avenues to pursue research concern identifying the specific actions that build trust among state-actors. Follow-on studies could consider the implementation of trust-building actions, for example, to reduce polarisation, sovereignty and/ or internet

fragmentation. Finally, examining trust-building measures as they play out in supporting the actions of entrepreneurs and norms on policy formation.

Should these findings prompt other researcher's interest future work could explore the nature of the trust amongst the policymakers involved in setting the agenda, and the role of citizens, in terms of trusting policy and/ or policy adoption. This would help policymakers understand the likely make-up and take-up of policy and to assess whether pilot testing or amendments may be necessary.

Acknowledgements

Thank you to university and industry colleagues for helpful discussions, challenging thoughts and help with the final version of the paper.

References

Accessnow. (n.d.) "Supreme court of India rules to restrict access to the worlds largest digital identity framework Aadhaar", [online]. https://www.accessnow.org/supreme-court-of-india-rules-to-restrict-worlds-largest-digital-identity-framework-aadhaar-but-debate-continues/ [Accessed 08 Jan. 2023].

Assembly, U.G. (2015) "Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security. Seventieth Session, Item, 93", [online]. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N15/228/35/PDF/N1522835.pdf [Accessed 08 Jan. 2023].

Bachmann, R. and Inkpen, A. C. (2011) "Understanding Institutional-based Trust-building Processes in Inter-organizational Relationships", Organization Studies 32, pp. 281–301.

Broeders, D. (2021) "The (im)possibilities of addressing election interference and the public core of the internet in the UN GGE and OEWG: a mid-process assessment", Journal of Cyber Policy, 6(3), pp. 277-297. DOI: 10.1080/23738871.2021.1916976

CA. 2021. "The United Nations, Our common Agenda", [online]. https://www.un.org/en/content/common-agenda-report/assets/pdf/Common_Agenda_Report_English.pdf [Accessed 08 Jan. 2023].

Claessen, E. (2020) "Reshaping the internet - the impact of the securitisation of internet infrastructure on approaches to internet governance: the case of Russia and the EU", Journal of Cyber Policy, 5(1), 140-157. DOI: 10.1080/23738871.2020.1728356

Currie, K.E. (2023) "Prevent China from killing human rights at the UN", Foreign Policy, 247, pp. 5–7, [online]. https://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=161154883&site=ehost-live&scope=site [Accessed 17 Feb. 2023].

Deutsch, M. (1958) "Trust and suspicion", The Journal of Conflict Resolution, 2(4), pp. 265-79.

digitalpublicgoods.net. 2022 "Promoting Digital Public Goods to promote a more inclusive world", [online]. https://digitalpublicgoods.net/ [Accessed 08 Jan. 2023].

Epifanova, E. (2020) "Deciphering Russia's sovereign internet law", German Council on Foreign Relations 16/Jan/2020, [online]. https://dgap.org/en/research/publications/deciphering-russias-sovereign-internet-law [Accessed 17 Feb. 2023].

EU (2012) "European Parliament resolution 12/Jun/2012 on critical information infrastructure

protection-achievements and next steps: towards global cyber-security" (2011/2284(INI)) https://www.europarl.europa.eu/doceo/document/TA-7-2012-0237_EN.pdf [Accessed 17 Feb. 2023].

Fulmer, A. and Gelfand, M. (2012) "At what level (and in whom) we trust: trust across multiple organizational levels", Journal of Management, 38(4), pp. 1167-1230.

Fukuyama, F. (1996) "Trust: The social virtues and the creation of prosperity", New York: Simon and Schuster.

Henriksen, A. (2019) "The end of the road for the UN GGE process: The future regulation of cyberspace", Journal of Cybersecurity, 5(1), 2019, tyy009, [online]. https://doi.org/10.1093/cybsec/tyy009 [Accessed 17 Feb. 2023].

hlpf.un.org. (2022) "Preserving an open Internet to achieve the SDGs: The Internet as an enabler for achieving SDGs 4 and 5", [online]. https://hlpf.un.org/2022/programme/preserving-an-open-internet-to-achieve-the-sdgs-the-internet-as-an-enabler-for [Accessed 08 Jan. 2023].

Lewicki, R.J., McAllister, D.J. and Bies, R.J. (1998). Trust and distrust: New relationships and realities. Academy of management Review, 23(3), pp.438-458.

Mayer R., Davis, J. and Schoorman, F. (1995) "An integrative model of organizational trust", Academy of Management Review, 20(3), pp. 709-734.

McKnight, D.H., Carter, M., Thatcher, J.B. and Clay, P. (2011) "Trust in a specific technology: an investigation of its components and measures", ACM Transactions on management information systems, 2(2), pp. 1-25.

Mir, U. B., Kar, A. K., Dwivedi, Y. G., Gupta, M. P. and Sharma, R. S. (2020) "Realizing digital identity in government: Prioritizing design and implementation objectives for Aadhaar in India", Government Information Quarterly, 37(2), [online]. 101442. doi.org/10.1016/j.giq.2019.101442 [Accessed 08 Jan. 2023].

O'Connor, S., Hanson, F., Currey, E. and Beattie, T. (2020) "Cyber-enabled foreign interference in elections and referendums", Canberra: Australian Strategic Policy Institute.

Oomsels, P. and Bouckaert, G., 2014. “Studying interorganizational trust in public administration: A conceptual and analytical framework for" administrational trust", Public Performance & Management Review, 37(4), pp. 577-604.

Rousseau, D.M., Sitkin, S.B., Curt, R.S. and Camerer, C. (1998) "Not so different at all: a cross discipline view of trust", Academy of Management Review, 23(3), pp. 393-404.

SDG.IISD.ORG (2021) SDG knowledge Hub, [online]. https://sdg.iisd.org [Accessed 08 Jan. 2023].

Russian State Duma. (2019) “On the Introduction of Changes into the Law on Information, Information Technologies and on the Protection of Information”, in Claessen, E. (2020) “Reshaping the internet - the impact of the securitisation of internet infrastructure on approaches to internet governance: the case of Russia and the EU”, Journal of Cyber Policy, 5(1), 140-157. DOI: 10.1080/23738871.2020.1728356

Schmidt, M. N. (2013) “Tallinn Manual on the International Law Applicable to Cyber Warfare”, New York: Cambridge University Press.

The Guardian. (2019) "Tim Berners Lee on 30 years of the web", [online]. https://www.theguardian.com/technology/2019/mar/12/tim-berners-lee-on-30-years-of-the-web-if-we-dream-a-little-we-can-get-the-web-we-want [Accessed 08 Jan. 2023].

The Register. (2022) "Indian authorities conflicting Aadhaar advice", [online]. https://www.theregister.com/2022/05/30/indian_authorities_conflicting_aadhaar_advice/ [Accessed 08 Jan. 2023].

UN, CA, 2021. “United Nations, Our Common Agenda”, [online]. https://www.un.org/en/common-agenda [Accessed 08 Jan. 2023].

UN, GDC. 2022. "The United Nations, Tech Envoy's Office, Global Digital Compact", [online]. https://www.un.org/techenvoy/global-digital-compact [Accessed 08 Jan. 2023].

UN, ODC. (2022a) “Cybercrime ad hoc committee, no.3, Statement by the Russian Federation”, 29/Aug.-09/Sept./2022, [online].

https://www.unodc.org/documents/Cybercrime/AdHocCommittee/Third_session/Documents/Statements/Item6/Russian_Federation_6_E.pdf [Accessed 18 Feb. 2023].

UN, ODC. (2022b) “Cybercrime ad hoc committee, no.3, Main session”, 29/Aug.-09/Sept./2022, [online]. https://www.unodc.org/unodc/en/cybercrime/ad_hoc_committee/ahc_third_session/main.html [Accessed 18 Feb. 2023].

UN, RDC. (2020) "The United Nations, GA Roadmap for digital cooperation: implementation of the recommendations of the high-level panel on digital cooperation", [online]. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/102/51/PDF/N2010251.pdf?OpenElement [Accessed 08 Jan. 2023].

UN, SDG. (n.d.) "The United Nations, Sustainable Development goals", [online]. https://sdgs.un.org/goals [Accessed 08 Jan. 2023].

UN, SG Panel. (2019) "UN Secretary General's High-Level Panel on Digital Cooperation", [online]. https://www.un.org/en/sg-digital-cooperation-panel [Accessed 08 Jan. 2023].

UN, Tech Envoy. (2022) "Contribute to the Global Digital Compact: A How-to-Guide (version 17 October 2022)", [online]. https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/Global-Digital-Compact_how-to-engage-guide.pdf [Accessed 08 Jan. 2023].

Vanneste, B.S. (2016) "From interpersonal to interorganizational trust: the role of reciprocity", Journal of Trust Research, 6(1), pp. 7-36.

Wylde A. (2022) “Cyber Security Norms: Trust and Cooperation,” Conference paper. ECCWS 2022.

Wylde, A. (2022) “Questions of trust in norms of zero trust”. In Intelligent Computing, Proceedings of the 2022 Computing Conference, 3, pp. 837-846. Cham: Springer International Publishing.

PhD Research Papers

Cybersecurity in Mozambique: Status and Challenges

Martina Jennifer Zucule de Barros[1] and Brett Van Niekerk[2]
[1]Universidade Pedagógica de Maputo, Maputo, Mozambique
[2]Durban University of Technology, KwaZulu-Natal, South Africa

martina.barros02@googlemail.com
brettvn@gmail.com

Abstract: Digital technologies became one of the most important components of societies day to day life. In Africa, they brought several benefits as well as challenges. For instance, the number of cyber-crimes and cyber-attacks are increasing. Yet, not all 54 African countries have implemented proper cybersecurity measures such as the adoption of national cybersecurity strategy, technical and organizational measures, development of cyber capacity and fostering national and engaging in regional and international cooperation. However, the adoption of these measures are vital and imperative. Mozambique is one of these countries where these measures are lacking. Therefore, the aim of this paper is to give an overview of the current state of cybersecurity in Mozambique considering all of the aspects mentioned above. Additionally, this paper also aims to present some best practices that Mozambique can adopt to improve and intensify its cybersecurity commitments. The proposed recommendations are based on internationally recognized frameworks and models developed by entities such as the European Union Agency for Cybersecurity (ENISA), International Telecommunication Union (ITU) and African Union (AU).

Keywords: cybersecurity, cyber- attacks, policy, standards, Mozambique

1. Introduction

Over the last years, developing countries, especially African countries have been experiencing a growth of their country's economy and social development due to their increasing reliance on the use of information and communications technologies (ICT) and cyberspace. These are considered as indispensable elements for governments, business and individuals and are viewed as crucial infrastructures of the 21st century. Despite all the benefits, they also bring several challenges. For instance, the number of cyber-attacks and cyber-crime is increasing globally. In Africa, Saeed and Osakwe (2021) state that many countries have been experiencing a rise of digital threats and malicious cyber activities for instance, sabotage of public infrastructure, national security breaches involving cyber-espionage and intelligence theft by militant groups, losses from digital fraud and illicit financial flows. Therefore, addressing these vulnerabilities requires a higher commitment to cybersecurity. This "requires enforceable policy safeguard, risk prevention and management approaches, along with technologies and infrastructure that can protect each country's cyber environment, as well as individual and corporate end-user assets" (Saeed & Osakwe, 2021). However, according to ITU's 2020 Global Cybersecurity Index (GCI) report, the level of African countries' commitments to cybersecurity remains low compared to other countries in other regions around the world (ITU, 2020).

In many African countries, cybersecurity is not viewed as a national priority as in developed countries. This is emphasized by Kshetri (2019) which states that in Africa, cybersecurity is considered to be a luxury and not a necessity. On the one hand, countries such as Mauritius and Tanzania stand out as regional leaders in cybersecurity. On the other hand, countries such as Mozambique, Lesotho and Madagascar continue to present lower levels of cybersecurity commitments (ITU, 2020). In Mozambique, cybersecurity has not yet been recognized by the government as a national priority. Only in 2021, the government of Mozambique approved the country's first national cybersecurity policy and the national cybersecurity strategy. Recently, the government of Mozambique officially published the policy (INTIC, 2021). Despite this there is no information regarding when the policy will come into effect i.e, will be implemented. In February 2022, a group of hackers attacked more than 30 web sites from the government of Mozambique.

According to the National Institute of Electronic Government (INAGE) despite the attacks, the problem was solved and no data breach was registered as well as loss of information. However, this was the first time that the country registered a cyber-attacks of this magnitude (Club of Mozambique, 2022). In the last three years, the Covid-19 pandemic has forced many people in developed as well as in developing countries to work remotely. This also contributed to an increasing number of cyber-attacks and has been considered as one of the world's biggest cybersecurity threats (Alawida et al., 2022). Therefore, this paper presents an overview of the cybersecurity environment in Mozambique. The paper addresses all issues related to the five pillars of the ITU global cybersecurity agenda (GCA). Frameworks such as the cybersecurity maturity model (CMM), the ITU global cybersecurity agenda (GCA) and the Cyber Readiness Index 2.0 were utilized to conduct this research. The rest

of this paper is organized as follows: Section 2 describes the cybersecurity environment in Mozambique. In section 3, an analysis of the state of cybersecurity is presented. The recommendations are outlined in section 4 and finally section 5 concludes the paper.

2. Cybersecurity environment in Mozambique

The ITU states that cybersecurity is built upon five strategic pillars such as: legal, technical, organizational measures, capacity development and international cooperation (ITU, 2017). These pillars measure countries' commitments to cybersecurity. Thus, this section discusses the state of these pillars in Mozambique.

2.1 Legal measures

In Mozambique, the government does not view cybersecurity as a critical issue. Indeed, only in 2016 the Mozambican government started to address issues related to cybersecurity with the development of the country's first national cybersecurity strategy draft. The process was conducted by the National Communications Institute of Mozambique (INCM) and supported by the ITU and Europe Union (EU). Also in 2016, the first national cybersecurity strategy draft was released. The document states that Mozambique aims to be "a nation with a safe, secure, and resilient cyberspace that enables a knowledge based and digital society economy" (INCM, 2016). However, this strategy has never come into effect.

In 2017, the Mozambican assembly approved the country's first Electronic Transaction Law (GovMZ, 2017). Also by 2017, the government of Mozambique designated the INTIC as the new entity to oversee all the cybersecurity issues at national level. In 2020, the country's first national cybersecurity policy and the national cybersecurity strategy were approved by the Mozambican government and in 2021, both documents were officially published (INTIC, 2021). Beyond that there is no other legislation or regulation related to cybersecurity. For instance, there is no specific regulation to deal with, data protection, cyber-terrorism, cyber-activism, cyber-espionage and cyber threats targeting national security. Furthermore, Mozambique does not have any mechanisms related to child online protection. In 2018, the Mozambican government ratified the AU convention on cybersecurity and data protection (Club of Mozambique, 2018). However, aspects related to cybersecurity and data protection are still not being incorporated in national laws. In addition, there is no awareness of cybersecurity standards and regulations in Mozambique.

2.2 Technical measures

In this pillar, Mozambique also lacks comprehensive plans. Despite this, institutions such as INGAE and the Mozambique research and education networks (MoRENET) have established their computer security incident response teams (CSIRTs). The CSRITs from INAGE covers the government institutions (INAGE, 2021) whereas the CSIRT from Mozambique research and education networks (MoRENET) covers the academic institutions i.e., public and private (MoRENET, n.d.). However, these centers need to be strengthened to be able to give appropriate responses. In 2021, the INAGE announced that the Mozambican government will establish an electronic platform to track cyberterrorists. The aim of this platform is to protect the websites of state institutions and other entities from cyber-attacks. According to INAGE, digital service providers and other operators of essential services should notify the Mozambican entities in charge of cybersecurity issues about cybersecurity incidents. (Club of Mozambique, 2021). It is also not clear if the country has effective cybersecurity monitoring measures to control its critical national information infrastructures. For instance, a study conducted by Vumo et al (2018), shows that several public and private websites in Mozambique do not follow basic security measures. The majority of them are not audited and many people responsible for managing them lack knowledge regarding the security aspects (Vumo et al., 2018).

2.3 Organizational measures

According to ITU, this pillar ensures that "cybersecurity is sustained at the highest level of the executive and assigning relevant roles and responsibilities to various national entities, and making them accountable for the national cybersecurity posture" (ITU, 2020). In Mozambique, the first entity overseeing national cybersecurity issues was the INCM (Symantec, 2016). However, in 2017, the government of Mozambique designated the INTIC as the new entity in the government responsible for overseeing all cybersecurity issues (INTIC, 2021). According to ENISA (2016), there are three types of governance structures to govern cybersecurity: centralized approach, decentralized approach and institutional cooperation with the private sector such as public-private partnerships. In Mozambique, it is not clear which structure is being implemented. On one hand, INTIC, INAGE, INCM and

MoRENET continue to work on issues related to cybersecurity independently. On the other hand, there is no clear cooperation among them.

2.4 Cybersecurity capacity development

This area is considered a crucial factor in cybersecurity. The ITU states that "securing the cyber domain through cybersecurity capacity building activities is key as it contributes to reducing issues such as digital divide and cyber risks" (ITU, 2020). In Mozambique there is a lack of plans and programs in this regard. A study conducted by Symantec (2016) states that there is a lack of awareness raising campaigns as well as development of digital literacy and skills among the public in Mozambique. Primary and secondary schools do not have e-safety programs promoting good practices. Nevertheless, the recent cybersecurity policy approved by the Mozambican government mentions cybersecurity capacity building, research and innovation as one of its key principles (INTIC, 2021). The Ministry of Science and Technology and Higher Education in Mozambique has neither developed nor launched a program or initiatives to encourage cybersecurity education and applied research at academic institutions as well as cybersecurity awareness in all parts of society. To date, the only available information related to the existence of courses related to cybersecurity is related to one private academic institution, the Higher Institute of Transport and Communications (ISUTC) which offers a course related to cybersecurity focusing on firewalls. This course is offered in partnership with the Palo Alto Networks Cybersecurity Academy (ISUTC, 2020).

2.5 Cooperation

In relation to international cooperation, it is not clear if the government of Mozambique considers cybersecurity as a top issue of its foreign policy or international negotiations. There is a lack of information related to the engagement of Mozambique in fostering international cooperation. However, international cooperation is viewed as an important factor in cybersecurity. According to ITU (2018), establishment of national and international cooperation is very important "to facilitate a constructive dialogue, developing trust and cooperation mechanisms, finding mutually acceptable solutions to common challenges, and creating a global culture of cybersecurity" (ITU, 2018). Despite this in 2018, the government of Mozambique ratified the AU convention on cybersecurity and data protection (Club of Mozambique, 2018). Recently, the government of Mozambique announced a partnership with Brazil to establish a national certification authority i.e., Mozambique Digital Certification System. According to the Ministry of Science and Technology and Higher Education this system will allow the integration of the public and private sector such as registration and identification systems of the members of the academic and scientific community in Mozambique through the use of emerging technologies for electronic identity (MCETS, 2022).

3. Findings and Discussion

The ITU states that cybersecurity is built upon five pillars such as: legal, technical, organizational measures, capacity building and international cooperation (ITU, 2017). From the analysis of the cybersecurity environment in Mozambique one can see that the country's cybersecurity commitments are still at its start-up stage. In 2017, Mozambique was classified as an initiating stage i.e., a country that has started to make commitments in cybersecurity and ranked one hundred and nine globally (ITU, 2017).

In 2018, the overall country commitments remained the same as in 2017 (ITU, 2018). Recently, the ITU published an updated version of its GCI, the 2020 ITU's GCI report. This report states that Mozambique presents two areas of relative strength: technical and legal measures (ITU, 2020). However, technical and legal measures present elements such as data protection regulation, critical infrastructure regulations, active CIRTs, regional cooperation and child online protection mechanisms. South Africas as well as Tanzania have enacted legislation related to cyber crime. To date, Mozambique has not enacted any of them. Furthermore there is no official information covering how the government of Mozambique will act to establish them. For instance, despite the existence of national CIRTs, there is a lack of information related to activities that are being carried out by these centers. Nevertheless, Hathaway et al (2015) state that "a well established national CSIRT provides reactive services above all else...the ability to respond to incidents by containing and mitigating incidents as they occur" (Hathaway et al., 2015).

In relation to national cybersecurity policy and strategy, the government of Mozambique approved in 2021, the country's first national cybersecurity policy and its implementation strategy. From the capacity building dimension, the score is low. There is no information related to the development of a national plan for growth

science and innovation in cybersecurity. In addition, there is a lack of investments in cybersecurity basic and applied research and development (R&D). In relation to cyber awareness initiatives there is also a lack of these initiatives. To date, there is no information related to the establishment, for instance, of a national campaign related to cybersecurity targeting the general public, business as well as the government. For instance, South Africa is one of the African countries participating in the Safer Internet Day (SID). SID is an awareness program from the EU which aims to make the internet safer, particularly for young people (Bada et al., 2018). Mozambique is one the countries in the sub-saharan which is still not participating in this campaign. The development of national cybersecurity industries is also lacking in Mozambique. As well as other developing countries, Mozambique heavily relies on the information technologies (IT) products developed by developed countries.

In relation to international cooperation, the government of Mozambique has not defined cybersecurity as a part of its foreign policy. Moreover, Mozambique is not participating in many international forums related to cybersecurity, for instance the Forum of Incident Response and Security Teams (FIRST), on the contrary to its neighboring countries such as South Africa, Zambia and Tanzania, Mozambique is not a member of this forum. Nevertheless, Mozambique is one of the few African countries who have ratified the AU convention on cyber security and personal data protection. Figure 1 below, shows the country's current score of cybersecurity commitments according to ITU's 2020 GCI report (ITU, 2020).

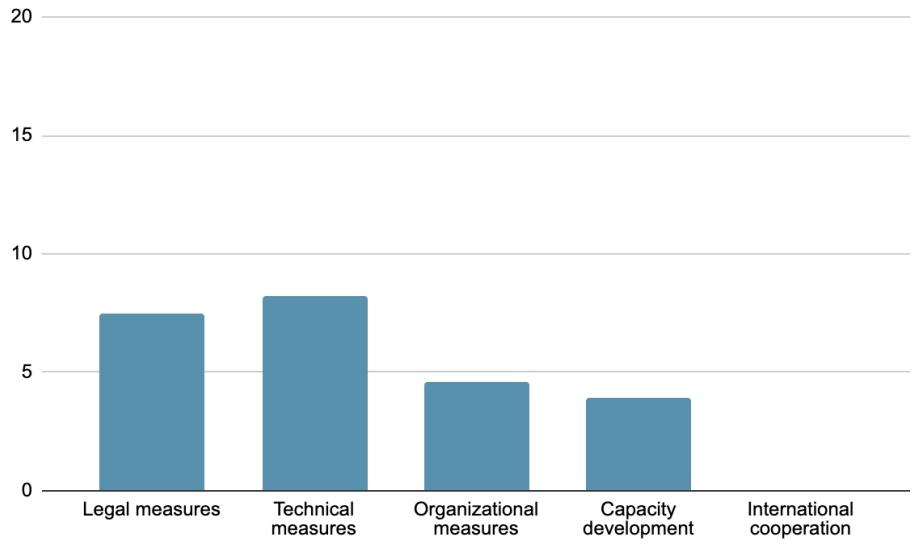


Figure 1: Mozambique's cybersecurity commitments level (ITU, 2020)

The score starts from 0 to 20 points (0.200) in each pillar. Thus, Mozambique's current overall score is 24.18 points. The GCI is the result of the total score of all five pillars and starting from 0 to 100 points. According to ITU, countries are ranked according to their final score. Therefore, Mozambique ranks 123 globally and regionally ranks 23 (ITU, 2020). On the other hand, the e-governance academy (eGA) states that Mozambique's national cybersecurity index (NCSI) is 9.09 points from (0 to 100 points) and ranks 151 globally. The NCSI "is a global live index, which measures the preparedness of countries to prevent cyber threats and manage cyber incidents" (e-GA, 2018). This index focuses on cybersecurity elements such as existing legislation, established units, cooperation formats and outcomes implemented by local governments.

4. Recommendations

After the analysis of the country current situation and based on our findings we proposed that, the government of Mozambique should:

- Enhance governmental coordination: Governments play a crucial role in managing cybersecurity issues. The ITU guide on developing a national cybersecurity strategy states that "cybersecurity should be promoted and sustained at the highest levels of government" (ITU, 2018). On the other hand, the recently published AU digital transformation strategy recommends member states to establish national cybersecurity governance structures involving actors such as policy makers,

technical, economic, educational and business communities, legal, law enforcement, military, academia and diplomacy (AU, 2020). Therefore, Mozambique should define cybersecurity as a national priority and improve its national governance structure.

- Implement a national cybersecurity strategy: Hathaway et al (2015) states that a national cybersecurity strategy should not only be articulated, it should also be actionable. Therefore, Mozambique should implement its recently approved national cybersecurity policy.
- Establish cyber response capabilities: The existence of national entities to exchange threat and vulnerability assessment and address operational cybersecurity challenges is very important (ITU, 2020). Moreover, cyber incident response capabilities play an important role in coordinating incident management at national level (ENISA, 2016). Therefore, Mozambique should establish national cyber incident response capabilities to oversee these issues.
- Strengthen public-private partnership (PPP): PPP plays a crucial role in cybersecurity efforts "from sharing actionable intelligence, exchanging good practice and communicating R&D needs and priorities" (ITU, 2020). Therefore, the government of Mozambique should establish effective partnership with other public and private entities.
- Invest in cybersecurity capacity building: The development of a stronger cybersecurity workforce is considered a crucial issue to decrease risks for business and society in general. According to ENISA (2016), increasing the level of skills and knowledge is an indispensable element in building society's resilience to cyberspace threats. Moreover, the AU digital transformation strategy recommends member states to conduct capacity building of policy makers and law enforcement to strengthen cybersecurity (AU, 2020). Thus, Mozambique should: i) develop a national plan covering cybersecurity across primary and secondary schools, ii) integrate cybersecurity courses in computer science, IT programmes and other related programmes, iii) develop dedicated undergraduate and graduate post-graduate programs in cybersecurity and iv) invest in research in cybersecurity.
- Implement awareness raising campaigns: Raising awareness is very important because "end users play a crucial role in ensuring the security of networks and information systems" (EU, 2013). Thus "they need to be made aware of the risks they face online and be empowered to take simple steps to guard against them" (EU, 2013). South Africa has been carrying out several awareness initiatives developed by academic institutions as well as the public and private sector. The University of South Africa (UNISA) has developed a cyber awareness campaign to contribute towards the creation of a cybersecurity culture targeting school children. The University of Pretoria (UP) runs a project called PumaScope targeting rural schools, churches and orphan homes (Dlamini & Modise, 2012). Therefore, Mozambique should develop and implement national cybersecurity awareness campaigns targeting the general public, business and the government.
- Develop and align cybersecurity legislation and regulation: The need for strong and effective legislation for instance, to tackle cybercrime is crucial. The AU's digital strategy recommends the development and adoption of regulations and legislation related to personal data protection, cybercrime, cybersecurity standards and governance (AU, 2020). Mozambique should adopt or create laws in the area of cybercrime, child online protection and data protection. Moreover, Mozambique should intensify activities aimed at implementing standards related to cybersecurity.
- Engage in international cooperation: Cybersecurity plays a key role in many areas of international relations such as human rights, economic trade, security, stability, peace and conflict resolution, commerce and arms controls (ITU, 2018). Mozambique should develop and improve its national, regional and international cooperation to foster knowledge exchange and learning.

5. Conclusion

African economies as well as other economies around the world are facing increasing cyber-attacks. This underscores the need to strengthen cybersecurity measures as well as investment in human capital. However, in Africa, on the contrary to developed countries, many countries are still lagging behind. Therefore, this paper has examined the cybersecurity environment in a developing country, in the SADC region, Mozambique. Data was collected from international organizations such as the ITU - GCI, the e-GA - NCSI and the Symantec report on cyber crime and cyber security trends in Africa. The findings show that despite some efforts, cybersecurity is still a challenging process in Mozambique. To date, Mozambique still presents lower levels of cybersecurity commitments in almost all the five pillars of the GCA. However, with the upsurge of cyber-attacks, the need to adopt measures to deal with it became a priority. Therefore, the paper made some recommendations for the

government of Mozambique in order to improve the country's current scenario. With these recommendations we believe that Mozambique will gradually start making positive signs to improve its cybersecurity commitments. On the other hand, this can also help other African countries in the SADC region as well as in other regions on the continent.

References

Alawida, M., Omolara, A. E., Abiodun, O., & Al-Rajab, M. (2022, August 11). A deeper look into cybersecurity issues in the wake of Covid-19: A survey. *ScienceDirect*, (34), 8176-8206.

AU. (2020, May 18). *The Digital Transformation Strategy for Africa (2020-2030)*. Retrieved August 23, 2022, from https://au.int/en/documents/20200518/digital-transformation-strategy-africa-2020-2030

Bada, M., Von Soms, B., & Agrafiotis, I. (2018). Reviewing National Cybersecurity Awareness in Africa: An Empirical Study. *repository.cam*, 78-83. htpps://doi.org/10.17863/CAM.40856

Club of Mozambique. (2018, July 25). *Mozambican government approves resolution on cybersecurity*. Club of Mozambique. Retrieved October 06, 2022, from https://clubofmozambique.com/news/mozambican-government-approves-resolution-on-cybersecurity/

Club of Mozambique. (2022, February 21). *Mozambique: State portals are back online after cyberattack- Noticias*. Club of Mozambique. Retrieved September 19, 2022, from https://clubofmozambique.com/news/mozambique-state-portals-are-back-online-after-cyberattack-noticias-210067/

Dlamini, Z., & Modise, M. (2012). *Cyber Security Awareness Initiatives in South Africa: A Synergy Approach*. Proceedings of ICIW.

e-GA. (2018). *National Cyber Security Index*. https://ncsi.ega.ee.

ENISA. (2016, November 14). *NCSS Good Practice Guide*. Retrieved July 20, 2022, from https://www.enisa.europa.eu/publications/ncss-good-practice-guide

EU. (2013, February 07). *Cybersecurity Strategy of the European Union: An open, Safe and Secure Cyberspace*. EU. Retrieved August 24, 2022, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52013JC0001

GovMZ. (2017). *Lei de Transacoes Electronicas*. Retrieved September 19, 2022, from https://www.portaldogoverno.gov.mz/por/Media/Files/Lei-de-Transaccoes-Electronicas

Hathaway, M., Demchak, C., McArdle, J., & Spidalieri, F. (2015, November). *Cyber Readiness Index 2.0*. Retrieved August 15, 2022, from https://www.potomacinstitute.org/academic-centers/cyber-readiness-index

INAGE. (2021). *CSIRT GOV*. Retrieved September 19, 2022, from https://csirt.gov.mz/?page_id=3486#quem_somos

INCM. (2016). *1 Reuniao sobre Estrategia Nacional de Seguranca Cibernetica*. INCM. Retrieved September 19, 2022, from https://www.cibersegurança.org.mz/

INTIC. (2021, September 1). *Governo aprova Politica e Estrategia Nacional de Seguranca Cibernetica*. INTIC. Retrieved September 19, 2022, from https://www.intic.gov.mz/?p=979

ISUTC. (2020). *Academia Palo Alto Networks do ISUTC*. Retrieved September 05, 2022, from https://www.isutc.ac.mz/wp-content/uploads/2020/11/Certificações-Academia-Palo-Alto_ISUTC_20.pdf

ITU. (2017). *Global Cybersecurity Index*. ITU. Retrieved August 08, 2022, from https://www.itu.int/en/ITU-D/Cybersecurity/Pages/global-cybersecurity-index.aspx

ITU. (2018). *Global Cybersecurity Index*. Retrieved August 10, 2018, from https://www.itu.int/en/ITU-D/Cybersecurity/Pages/global-cybersecurity-index.aspx

ITU. (2020). *Global Cybersecurity Index 2020*. ITU. Retrieved August 11, 2022, from https://www.itu.int/en/ITU-D/Cybersecurity/Pages/global-cybersecurity-index.aspx

Kshetri, N. (2019). Cybercrime and Cybersecurity in Africa. *Journal of Global Information Technology*. 10.1080/1097198X.2019.1603527

MCETS. (2022). *Mocambique trabalha na operacionalizacao do SIstema de Certiicacao Digital*. Retrieved September 20, 2022, from https://www.mctes.gov.mz/mocambique-trabalha-na-operacionalizacao-do-sistema-de-certificacao-digital/

MoRENET. (n.d.). *who we are*. Retrieved September 10, 2022, from https://csirt.morenet.ac.mz/en/who-we-are/

Saeed, M., & Osakwe, S. (2021, September 1). *Are African countries doing enough to ensure cybersecurity and Internet safety*. ITU. Retrieved October 06, 2022, from https://www.itu.int/hub/2021/09/are-african-countries-doing-enough-to-ensure-cybersecurity-and-internet-safety/

Symantec. (2016, November). *cyber Crime & Cyber Security Trends in Africa*. Retrieved June 28, 2022, from https://thegfce.org/initiatives/cybersecurity-and-cybercrime-trends-in-africa/

Vumo, A., Spillner, J., & Koepsell, S. (2018). Analysis of Mozambican websites: How do they protect their users? *IEEE*, 8. 10.1109/ISSA.2017.8251780

Spreading Lies Through the Cyber Domain

Thomas Dempsey
American Military University, Charles Town, USA

thomas.dempsey@mycampus.apus.edu

Abstract: The expansion of Information Operations (IO) over the past ten years has allowed individuals and groups to increase their sphere of influence on a global scale. Nation-state cyber threat actors have increased their presence on social media, building out false personas to influence large populations. This type of activity is difficult to stop due to the availability of social networks on the internet and the ease of creating false personas that can't be directly attributed to the actor. IO activity has been observed with the Russian cyber activity during the 2016 U.S. Presidential elections and from Russian social media campaigns provoking extremist groups and attempting to cause physical harm, such as the 2017 campaign on Facebook to start a rally and a simultaneous counter rally in front of the Islamic Da'wah Centre of Houston. Although Russia has been observed leveraging this capability, they are not the only global actor in the cyber domain taking advantage of IO. Global threat actors have leveraged social media platforms and blogs to influence the global population and spread propaganda. This type of activity has been seen within traditional warfare using propaganda techniques. With the introduction of the cyber domain into warfare, there is an increased ability to communicate not only to one population but to the global community with the intent to manipulate the masses using IO. This paper examines the Cybersecurity Operations (CO) that have been observed utilizing IO and the psychological impacts they have had in successful campaigns against the United States. This paper argues that with increased influence capabilities in the cyber domain, individuals and groups will continue using IO to support tactical and strategic objectives. Through the available literature, this paper examines the impacts that IO has had on the United States through attempts to manipulate elections and create divides in the nation over the last ten years. This paper leverages the psychology of group processes to analyze the literature involving social media campaigns and the influencing of groups through the lens of social identity theory to provide new insight into mitigating and countering IO.

Keywords: Information Operations, Disinformation, Cybersecurity Operations, Social Identity, Cyber, psychology

1. Introduction:

This research examines the question of what are the psychological impacts of disinformation and Information Operation campaigns on the population of the United States between 2012 and 2022? Unlike other time periods in U.S. history, there has been a significant increase in disinformation and Information Operations (IO) over the past ten years affecting the U.S. population. These IO campaigns against the United States have attempted to impact the U.S. elections and have been attributed to countries such as Russia, China, and Iran (Martin et al. 2022). The campaigns have spread disinformation regarding political parties and attempted to influence public opinion either for or against specific candidates in support of the nation-state's national objectives. This has been seen with the Russian campaigns supporting Trump in the 2016 elections and the 2018 midterm elections (Martin et al. 2022), along with attacking Hillary Clinton during the 2016 elections (Martin et al. 2022). With the increase of disinformation and IO campaigns, it is becoming important to understand the impact and effectiveness of these campaigns. This research examines the hypothesis that successful disinformation or IO campaigns will have a psychological impact on the target audience by manipulating the social identity by reinforcing an existing social-political standpoint or changing the social-political standpoint of the target audience.

The United States National Institute of Standards and Technology (NIST) defines Information Operations as "The integrated employment, during military operations, of information-related capabilities in concert with other lines of operation to influence, disrupt, corrupt, or usurp the decision-making of adversaries and potential adversaries while protecting our own. Also called IO" (NIST 2022). In other words, Information operations are the use of information to influence an audience and can be used to shape perceptions and understanding of a targeted population to achieve the tactical or strategic goals of the executing party. Information operations have been around for a long time. During World War II, President Roosevelt established the U.S. Office of War Information (OWI), specifically tasked with information activity and propaganda in the U.S. and overseas (Roholl 2012). Altheide and Grimes (2005) argue that the Iraq war was a successful propaganda campaign in that the American public was not aware of the full context and that the storyline from the media was that Hussein was involved in the 9/11 attacks, he supported terrorism and planned to use Weapons of Mass Destruction (WMD) against the U.S. creating a sense of urgency to intervene.

Modern disinformation operations are achieved through social media platforms like Twitter and Facebook, along with other online forums like Reddit. With social media, there is a larger audience that can be accessed

throughout the internet, and information operations are easier to conduct. Social media has been growing exponentially, making it easier than ever to spread false information and have an impact on the population. During the 2022 U.S. midterm elections, social media platforms such as Telegram were used by Russian-associated hacktivist groups to claim Distributed Denial of Service (DDoS) attacks against political party websites and promote the narrative that Russia can influence the U.S. elections (Wahlstrom et al. 2022). The increased accessibility to information on the internet has made disinformation a significant problem in the information age. Information operations are not just about hacking or cyber warfare anymore. These operations use data to influence a target audience's emotions, beliefs, attitudes, and behaviours.

The United States is a target and instigator of disinformation campaigns that foreign actors carry out to support their national objectives. In a declassified 2021 Intelligence Community Assessment (ICA) on the foreign threats to the 2020 U.S. elections, it was identified that there were multiple information operations in the form of influence campaigns from foreign actors, including Russia, Iran, Lebanese Hizballah, Cuba, and Venezuela (National Intelligence Council 2021). These campaigns can have a profound psychological impact on the U.S. population. Due to the anonymity that can be achieved through social media, it can be difficult to track down those responsible for spreading disinformation. A greater understanding of the psychology behind the disinformation and its psychological impacts on the population is required to better prepare for the next attack. This paper will explore how the use of disinformation in information operations has become commonplace and what psychological impacts it has on people within the United States.

2. Literature Review:

The relevant literature regarding disinformation and information operations (IO) focuses on what is described as "fake news". Pathak et al. identify, "Disinformation intentionally attempts to mislead people into believing a manipulated narrative to cause chaos or, in extreme cases, violence." (Pathak et al. 2021). Pathak et al. argue that using manual processes, organizations cannot handle the speed required to verify or fact-check the volume of information being released and disseminated through the internet (Pathak et al. 2021). In the research, Pathak et al. show that through the use of natural language inference (NLI), there is promise in moving away from manual processes currently used for verifying fake news by implementing automated or semi-automated processes that can identify fake news and apply fact-checking capabilities (Pathak et al. 2021). Using automated processes to verify and fact-check fake news may be effective; however, there is no mechanism to identify the root source that distributed the content that would allow for attribution or identification of IOs. Rubin identifies that the ability to attribute news content to a source is critical when judging the content's credibility; with the news becoming increasingly available in decontextualized forms such as social media, it becomes even more difficult to identify the sources (Rubin 2019).

Tandoc et al. used a different method when evaluating fake news. They evaluated the characteristics of the fake news and what elements of the fake news were shared with legitimate news. The analysis from Tandoc et al. found, "The analysis showed that 98.6% of the articles analyzed included the news value of timeliness; 89.2% included the news value of negativity; 79.7% included the news value of prominence; but only 32% included the news value of impact." (Tandoc et al. 2021). In their study, it was also identified that topics about government and politics were among the highest percent of topics that consisted of fake news, with 51.6% of the analyzed news being related to these topics; terrorism or crime were the second highest topics, with 19.5% of content analyzed being related to these topics and science, health, and technology were in the lower percentages at around 10.3% of the content being related to these topics (Tandoc et al. 2021). Through the research of Paterson and Hanley, it is identified that within the United States, state security failed to detect the 2016 Russian interference with the Presidential election (Paterson and Hanley 2020). Paterson and Hanley state, "The use of social media as part of larger information warfare operations, combined with malicious programming and the promotion of political distrust, is a major challenge and one that democracies have not done enough to address." (Paterson and Hanley 2020).

In a disinformation study that was conducted by Raman et al. (2020), it was found that when individuals are presented with disinformation, the responses are still high enough to where a targeted attack utilizing disinformation could provoke a response that could result in negative effects on physical infrastructure (Raman et al. 2020). Raman et al. argue that based on the findings in their research, there need to be communication channels and mitigation strategies established by policymakers that address the identified vulnerability and would allow for warnings to be distributed to the public in an effort to reduce the responses to disinformation campaigns (Raman et al. 2020). Pathak et al. identified, "Since it is relatively easy and inexpensive to rapidly disseminate content through social media platforms, there is an urgent need for automated solutions to combat

disinformation." (Pathak et al. 2021). Along with the identified need for automated solutions to combat disinformation, Raman et al. displayed through their study that the response to disinformation is high enough that disinformation can be weaponized and may have the capability to cause physical damage (Raman et al. 2020).

For international campaigns, Lanoszka (2019) argues that disinformation campaigns with an international scope are ineffective at disrupting the balance of power and changing the decisions of the target state. Three primary barriers prevent disinformation from being effective internationally, uncertainty over international signals, pre-existing ideologies, and countermeasures that are in place to prevent disinformation (Lanoszka 2019). The reviewed literature did not address the potential impacts that disinformation campaigns have on the psychology of a population. Raman et al. show how disinformation and IO campaigns can be weaponized and used to conduct targeted attacks that invoke the target population to take the desired response action. Rubin identified that people only have a "slightly better than chance" of identifying deception and are generally highly susceptible to disinformation (Rubin 2019).

To expand on the existing research regarding disinformation and IO, psychology can be leveraged through the lens of social identity theory and potentially provide additional insight into the targeting of campaigns. There are three basic principles of social identity theory defined by Brown and Pehrson "First, people seek a positive self-concept (they want to see themselves in a positive light). Second, because the evaluation of groups that they belong to carries implications for this positive self-concept, individuals will try to achieve a positive social identity. Third, since evaluating the value of a social identity is essentially comparative, people prefer to see their groups as superior to other groups, at least on the criteria that matter most to them." (Brown and Pehrson 2019). From the definition of the social identity principles described by Brown and Pehrson (2019), in social identity theory, groups prefer to have an image of superiority; they will attempt to promote a positive social identity of themselves and within the group. Applying social identity theory to disinformation and IO campaigns may contend that social groups may have a greater likelihood of neglecting to verify disinformation that is in line with their beliefs, supports their group's social image, and provides a positive persona for the group. The neglect to verify information may cause the groups to inadvertently spread false information in efforts to establish supremacy for the group, which would increase the impact of a disinformation campaign.

The definition of disinformation provided by Pathak et al. is that it is false and misleading information that is shared intentionally to cause harm (Pathak et al. 2021). One of the primary objectives of disinformation is to impact and disrupt the balance of power (Lanoszka 2019). It is important to recognize that there is also a difference between disinformation and misinformation. Unlike disinformation, misinformation tends to have inaccuracies or errors and is typically unintentional (Rubin 2019). Verifying the source of information to verify its integrity has become increasingly difficult with the speed at which information is spread through social media platforms (Rubin 2019). Falls argues that disinformation is misleading information with the primary function of misleading (Fallis 2015). Disinformation is similar to lying, which creates different indicators to detect disinformation versus misinformation, where someone doesn't know what they are talking about and is unintentionally spreading information (Fallis 2015). Expanding on the existing research will provide insight into the psychological impacts and group processes that may cause disinformation to spread intentionally, increasing the reach of IO and disinformation campaigns.

3. Methods:

This research seeks to expand on the existing literature and identify the number of disinformation or IO campaigns that have taken place against the United States, and to identify any psychological impacts through the lens of social identity theory. According to the reviewed literature from Paterson and Hanley (2020), nation-state cyber actors are utilizing disinformation and IO to cause disruptions in the political processes of the United States, such as interference in the 2016 Presidential election. Are these targeted attacks successful at manipulating the political views of the nation? Raman et al. (2020) showed through their research that there was a high enough follow-through based on disinformation to cause blackouts of the power grid. The hypothesis for this research is that successful disinformation or IO campaigns will have a psychological impact on the target audience by manipulating the social identity by reinforcing an existing social-political standpoint or changing the social-political standpoint of the target audience. Examining identified campaigns, the research attempts to identify the psychological impacts on the social-political standpoints of the target audience of the campaign.

Utilizing the ten-year timeframe identified in the research question, correlations have been made using the election data. The analysis will focus on IO and disinformation that has targeted political campaigns. In an effort

to identify campaigns targeting political parties, multiple datasets have been examined. The datasets have been obtained using a combination of open-source data repositories that have collected known fake news from social media and online news sources. This known fake news and data regarding the mentions of information operations and disinformation will be correlated with data scrapes conducted from multiple social media platforms from 2016 through 2019 to identify how popular the known fake news relating to the political parties was through that timeframe. The political party polling data has been used to look at changes to the support for the political party aligned to the associated timeframe of the campaigns. From the three principles of social identity theory, this research examines how disinformation propagates through social groups and expands into successful disinformation campaigns.

Using a semi-quantitative approach and utilizing the data from the MIT Election Data and Science Lab, the election results were examined for the House, Senate, and Presidential elections from 2012 through 2021. During this timeframe, there were three different U.S. Presidents in office. From 2012 through 2017, President Obama finished his second term as President. From 2017 through 2021, President Trump served as President, and from 2021 to the time of this research, President Biden has served as the President. Based on the U.S. Presidents in office over the identified timeframe, two different political parties held office, President Obama 2012 through 2017 and President Biden 2021 to the present as the Democratic candidate, and President Trump 2017 through 2021 as the Republican candidate. This information was also used when collecting data from open-source datasets.

Nine keywords were used to search through the data to identify where there are mentions of political affiliations. The nine keywords used were House, Senate, Democrat, Republican, Obama, Trump, Hillary, Sanders, and Biden. These nine keywords were selected based on three being the names of former or current U.S. Presidents as identified earlier. Hillary and Sanders were used based on those being the popular names for the democratic candidates in the 2016 election. Democrat and Republican were chosen due to those being the names of the primary parties in the U.S. elections. Senate and House were used to represent the other areas of the U.S. government that the identified presidential candidates did not capture. The data available from the MIT Election Data and Science Lab for the Senate elections spans from 2012 through 2021. For the House elections, the MIT Election Data and Science Lab dataset only contained data from 2012 through 2020. To maintain consistency in examining the data, only the data from 2012 through 2020 was utilized from all three of the datasets. The identified IO and disinformation campaigns were aligned to the election results to highlight0 if there are any significant correlations between the campaigns and the election results data. A mixed-method research approach was taken to quantify the areas where disinformation has been observed. Based on the results from the semi-quantitative approach, a qualitative analysis will be conducted to examine the potential psychological impacts of the political standings observed in the United States over the examined timeframe.

4. Findings and Analysis:

Examining the dataset sourced from FakeNewsNet (Shu et al. 2017 – 2018), the data captures the number of tweets that mentioned the keywords within the dataset. Using the nine keywords House, Senate, Democrat, Republican, Obama, Trump, Hillary, Sanders, and Biden, the dataset was examined to identify how many tweets were associated with the identified fake news and the respective keyword. The results of the number of tweets associated with fake news by the keywords can be seen in Figure 1.

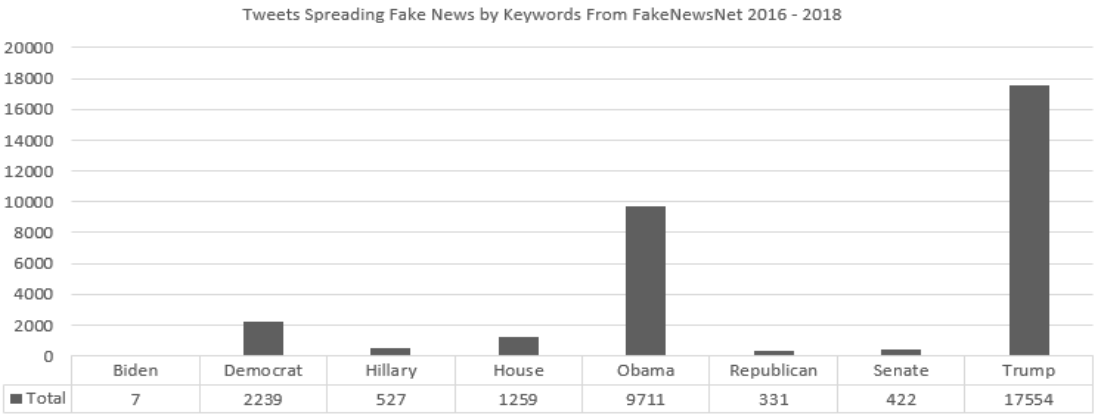


Figure 1: Number of Fake News Tweets About the Keywords (Shu et al. 2017 - 2018)

It is important to note that there were no relevant results in the dataset for the keyword Sanders and limited data regarding the keyword Biden. From the data relating to fake news, the number of results that related to the keyword Trump in 2016 through 2018 timeframe was a great deal higher than those relating to any of the other keywords. The second highest was Obama, with 9711 identified tweets associated with the keyword dealing with fake news. The timeframe of this dataset overlaps with the 2016 election when President Obama left office, and President Trump was coming into office. Martin et al. identified 20 information operations that took place against the United States, starting from 2014 through 2019. The country and timeframes are displayed in figure 2. During the 2014 attacks, the primary threat actor was Russia, and the three attacks aimed to spread misinformation, undermined the image of Obama, and polarize politics in America (Martin et al. 2022). The operations starting in 2015 also came primarily from Russia. They focused on attacking Hillary Clinton in the 2016 elections, attacking the democratic party during the 2016 and 2018 elections, discrediting institutions in America, and supporting Trump in the 2016 and 2020 elections (Martin et al. 2022). In the campaigns observed starting in 2016, there were additional threat actors, including Iran and unknown actors conducting information operations. These campaigns intended to attack Hillary Clinton (UNK), attack critics of Trump after the 2016 elections (RUS), support Trump in the 2016 election and 2018 midterm elections (UNK), support independence movement in California and Texas (RUS), attack Trump in the 2016 and 2020 elections (IRN), and support Alt-right movements in the 2016 and 2020 elections (RUS) (Martin 2022). In 2017 there were campaigns observed from China, Russia, and Iran with the intent to support Republican senator Roy Moore (RUS), attack the Republican party (IRN), and Polarize politics in America (CHN, IRN) (Martin 2022). In 2018 only Russia and Iran were observed with the intent to support Trump's Nominees in Courts (RUS) and promote Iranian foreign policy initiatives (IRN) (Martin 2022). The campaign starting in 2019 was from Russia and focused on attacking Biden in the 2020 election (Martin 2022).

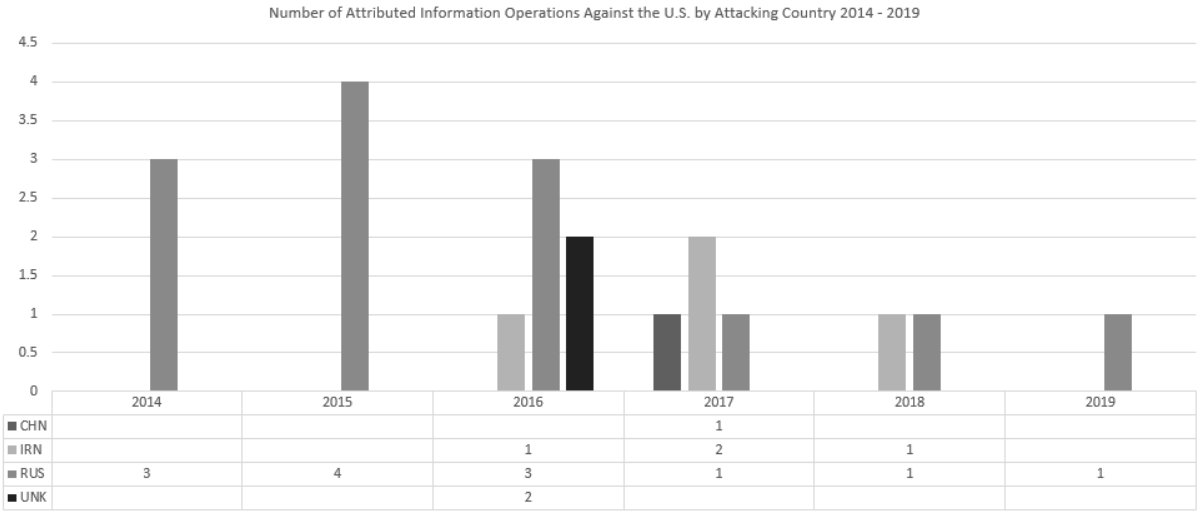
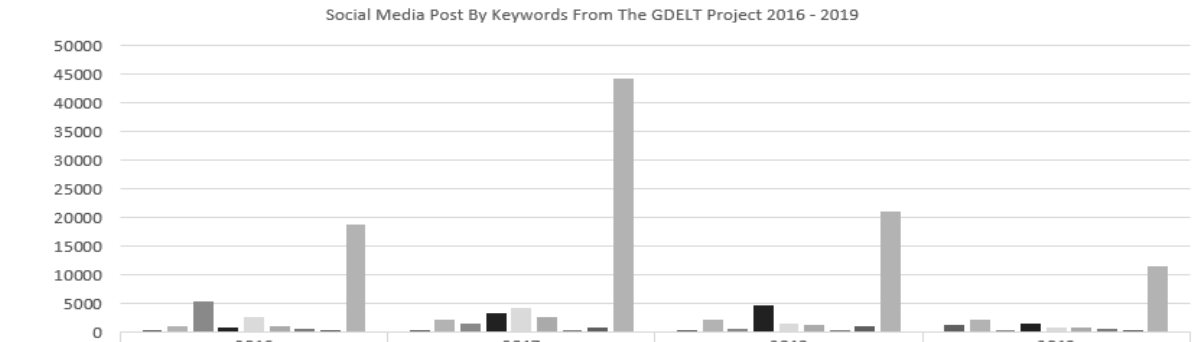


Figure 2: Information Operations Against the U.S. (Martin et al. 2022)

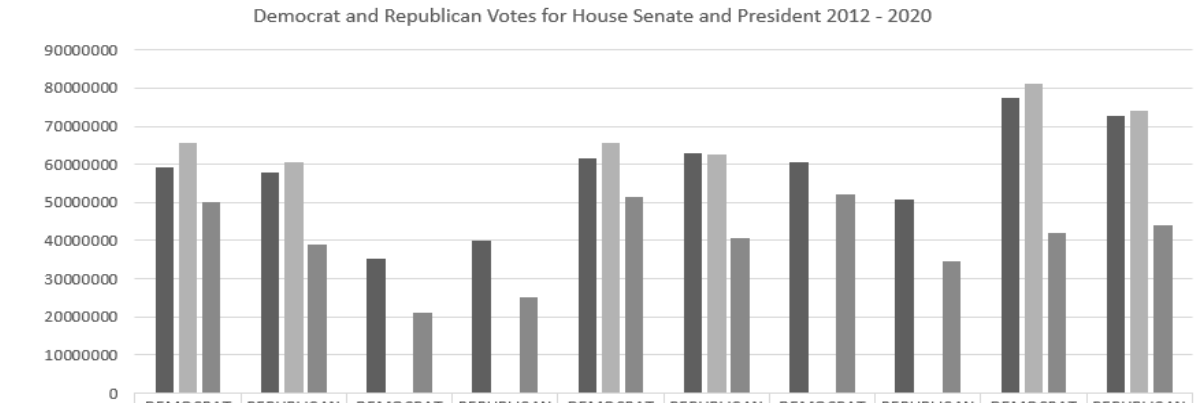
Due to the timeframe captured with the fake news dataset, there is no data to compare for the 2014 and 2015 Russian campaigns. There is data for 2016 – 2018; based on the keywords, there was fake news identified with the keywords Trump and Obama along with Hillary, House, and Democrat, which aligns with the majority of the identified campaign's intent from the 20 campaigns. To gain an idea of the social media landscape during that timeframe, additional social media posts were examined to understand how many mentions the keywords received from a general dataset. This was done by using the Global Database of Events, Language, and Tone (GDELT) Project dataset that captured social media posts from Facebook, Instagram, QQ, Twitter, Vimeo, VK, and YouTube from April 2016 through the end of September 2019 (The GDELT Project 2019). Figure 3 shows the breakdown of social media posts relating to the keywords during the timeframe captured in the dataset. When examining the dataset from GDELT, only posts that discussed the keywords were examined. Any post originating from an account associated with the keyword was not counted toward the keyword.



	2016	2017	2018	2019
Biden	203	342	239	1215
Democrat	1045	2123	2275	2126
Hillary	5395	1486	599	252
House	787	3245	4730	1407
Obama	2724	4306	1399	787
Republican	971	2605	1253	883
Sanders	588	370	408	607
Senate	316	910	942	290
Trump	18676	44236	21012	11511

Figure 3: Social Media Post by Keywords (The GDELT Project 2019)

From GDELT dataset, there is a consistently high number of social media posts involving the keyword Trump. Similar to what was observed in the fake news data, in 2017, Obama is second in the number of social media posts. Because this is looking at the social media post and not considering the number of shares or associated tweets such as the fake data, the number for Obama is lower than what is observed in the fake news dataset. The GDELT dataset showed the greatest numbers in 2017, and overall, the numbers appear to be declining for the keywords Trump, Republican, Obama, Hillary, and Biden moving into 2018. The GDELT dataset contained data through September 2019, meaning there is not a full year's worth of data for 2019. Though there is not a full year of data, there is still an increase in the mentions relating to the keyword Biden and Democrat. Figure 4 contains the voting results for elections that took place for the House, Senate, and Presidential elections from 2012 through 2020. What would be expected is that if the IO conducted during these timeframes were successful, there would be a clear deviation in the voting results showing the social-political impact of the campaigns.



	DEMOCRAT	REPUBLICAN	DEMOCRAT	REPUBLICAN	DEMOCRAT	REPUBLICAN	DEMOCRAT	REPUBLICAN	DEMOCRAT	REPUBLICAN
	2012		2014		2016		2018		2020	
US HOUSE	59194640	57977915	35273163	39926535	61417454	62772633	60454544	50752541	77336102	72671813
US PRESIDENT	65752017	60670117			65677288	62692670			81268908	74216146
US SENATE	49988282	39128301	21246466	25010851	51315969	40842211	52208408	34721533	42138692	43948213

Figure 4: U.S. Votes by Party (MIT Election Data and Science Lab 2017)

During 2016 when the highest number of IO campaigns were observed targeting the democratic candidate Hilary Clinton and supporting Donald Trump, these results are similar to the results observed in the 2012 elections. In the voting numbers from the 2016 elections, the democratic party won in both the Senate and the President, and the Republican candidate won the Presidential election through the electoral college, not from the overall

votes received. Based on the examined data, it does not appear that the observed IO campaigns met their intent or that the disinformation spread through fake news had a large enough impact on the targeted population to noticeably change the social-political views of the population.

5. Conclusion:

This research examines the question of what are the psychological impacts of disinformation and Information Operation campaigns on the United States between 2012 and 2022? The intent is to test the hypothesis that a successful disinformation or IO campaign will have a psychological impact on the target audience by manipulating the social identity by reinforcing an existing social-political standpoint or changing the social-political standpoint of the target audience. Based on the available data that was used in this research, it does not suggest that the observed IO and disinformation campaigns were successful at changing the social-political standpoint of the target audience. Further research is required to understand the scope of the impact that IO and disinformation may have on the psychology of a targeted population. Although Raman et al. (2020) displayed that disinformation can be weaponized to potentially cause physical damage if there are high enough responses to the campaign. The findings of this research further support Lanoszka's analysis of international disinformation campaigns being ineffective at changing the decisions of the target state and disrupting the balance of power (Lanoszka 2019). From the three primary barriers discussed by Lanoszka, the pre-existing ideologies (Lanoszka 2019) may be a contributing factor as to why the voting data showed similar trends from 2012 and 2016, even though there were more disinformation and IO campaigns taking place during the 2016 timeframe. Although the impact of the IO and disinformation campaigns didn't appear to be significant from the observed data, further research is required to focus on where the IO and disinformation campaigns have been effective. Using IO and disinformation has continued to be a relevant tactic deployed by nation-states, as observed from China, Russia, and Iran through this paper. It is important to gain a better understanding of what creates a successful campaign to enable safeguards to be deployed that can mitigate the impact of a successful campaign.

References:

Altheide, D. L., and Jennifer N. Grimes. (2005). War Programming: The Propaganda Project and the Iraq War. [online], The Sociological Quarterly, 46(4), 617–643. http://www.jstor.org/stable/4121509

Brown, R. and Pehrson, S. (2019) Group Processes (3rd Edition), [online], Wiley Global Research (STMS), https://online.vitalsource.com/books/9781118719312

Fallis, D. (2015) 'What Is Disinformation?', [online], Library Trends, 63(3), pp. 401–426. https://doi.org/10.1353/lib.2015.0014.

Lanoszka, A. (2019) 'Disinformation in international politics', [online], *European Journal of International Security*, 4(2), pp. 227–248. https://doi.org/10.1017/eis.2019.6.

Martin, D. A., Shapiro, J. N., & Ilhardt, J. G. (2022). Introducing the Online Political Influence Efforts dataset. [online], Journal of Peace Research, 0(0). https://doi-org.ezproxy1.apus.edu/10.1177/00223433221092815

Martin, D. A., Shapiro, J. N., & Ilhardt, J. G. (2022) "Replication data and Online Appendix for: "Introducing the Online Political Influence Efforts dataset" Journal of Peace Research", [online], Harvard Dataverse V1, https://doi.org/10.7910/DVN/8IF59Q,

MIT Election Data and Science Lab (2017) "U.S. House 1976–2020", [online], Harvard Dataverse, V11, UNF:6:ry6R0P1KRBhWkIfZzKiM8A== [fileUNF], https://doi.org/10.7910/DVN/IG0UN2

MIT Election Data and Science Lab (2017) "U.S. President 1976–2020", [online], Harvard Dataverse, V6; 1976-2020-president.tab [fileName], UNF:6:4KoNz9KgTkXy0ZBxJ9ZkOw== [fileUNF], https://doi.org/10.7910/DVN/42MVDX

MIT Election Data and Science Lab (2017) "U.S. Senate 1976–2020", [online], Harvard Dataverse, V6, UNF:6:dogvks8KPD0c/hzNi9kaag== [fileUNF], https://doi.org/10.7910/DVN/PEJ5QU

National Institute of Standards and Technology (NIST) (2022) "information operations (IO)", [online], Information Technology Laboratory, Computer Security Resource Center, https://csrc.nist.gov/glossary/term/information_operations

National Intelligence Council (2021) "Foreign Threats to the 2020 US Federal Elections", [online], Intelligence Community Assessment, https://www.dni.gov/files/ODNI/documents/assessments/ICA-declass-16MAR21.pdf

Paterson, T. and Hanley, L. (2020) "Political Warfare in the Digital Age: Cyber Subversion, Information Operations and 'Deep Fakes.'", [online], Australian Journal of International Affairs 74, no. 4: 439–54. https://doi.org/10.1080/10357718.2020.1734772.

Pathak, A., Rohini, S., and Nihit, N. (2021) "Disinformation: Analysis and Identification.", [online], Computational and Mathematical Organization Theory 27, no. 3: 357–75. https://doi.org/10.1007/s10588-021-09336-x.

Raman, G., AlShebli, B., Waniek, M., Rahwan, T., and Peng, J. (2020) "How Weaponizing Disinformation Can Bring down a City's Power Grid.", [online], Edited by Qi Jiang. PLOS ONE 15, no. 8: e0236517. https://doi.org/10.1371/journal.pone.0236517.

Roholl, M. (2012) "Preparing for Victory. The U.S. Office of War Information Overseas Branch's illustrated magazines in the Netherlands and the foundations for the American Century, 1944-1945", [online], European Journal of American Studies, vol. 7, no. 2.

Rubin, V. (2019) "Disinformation and Misinformation Triangle: A Conceptual Model for 'Fake News' Epidemic, Causal Factors and Interventions.", [online], Journal of Documentation 75, no. 5: 1013–34. https://doi.org/10.1108/JD-12-2018-0209.

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. (2018) "FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media" arXiv preprint arXiv:1809.01286

Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017) "Fake News Detection on Social Media: A Data Mining Perspective" ACM SIGKDD Explorations Newsletter, volume 19 no. 1 pg. 22-36.

Shu, K., Wang, S., and Liu, H. (2017) "Exploiting Tri-Relationship for Fake News Detection" arXiv preprint arXiv:1712.07709

Tandoc, E., Thomas, R., and Bishop L. (2021) "What Is (Fake) News? Analyzing News Values (and More) in Fake Stories", [online], Media and Communication 9, no. 1: 110–19. https://doi.org/10.17645/mac.v9i1.3331.

The GDELT Project (2019) "Compiling A Master List Of Social Media In The News 2016-2019", [online], https://blog.gdeltproject.org/compiling-a-master-list-of-social-media-in-the-news-2016-2019/

Wahlstrom, A., Xia, J., Revelli, A., and Serabian, R. (2022) "Information Operations Targeting 2022 U.S. Midterm Elections Include Trolling, Narratives Surrounding Specific Races, Politicians", [online], Mandiant, https://www.mandiant.com/resources/blog/information-operations-2022-midterm-elections

The Concept of Comprehensive Security as a Tool for Cyber Deterrence

Maria Keinonen
National Defence University, Helsinki, Finland

maria.keinonen@mil.fi

Abstract: Cyber deterrence is often studied from the point of view of deterrence by punishment or offensive cyber strategies. A vast amount of studies claim that deterrence in cyberspace can never be successful with cyber means alone due to technical challenges and the problem of attribution. Some scholars argue that cyber resilience is an essential part of cyber deterrence, since not every cyberattack can be countered. These reviews are usually technical and concentrate on investigating the balance of offensive and defensive cyber strategies. The technical view leaves gaps in the physical and cyber-persona layers of cyberspace. This paper examines resilience from a societal perspective and reflects on the findings of cyber deterrence theories. The Concept of Comprehensive Security (CCS) is a Finnish model for building and sustaining resilience in society. Preparation for disruptive situations is carried out with the operating principle of overall safety, where society´s vital functions are protected in collaboration between the authorities, the business world, organisations, and citizens. The growing importance of cyber security has led to emphasising the importance of cyber resilience in the Concept of Comprehensive Security. This study investigates the possibilities to utilize the CCS as a tool for cyber deterrence and aims to create a new perspective on the international academic discussion of cyber deterrence. The research method is content analysis. The investigated material consists of Finnish CCS documents, as well as academic cyber deterrence and cyber resilience literature. The characteristics of the CCS are compared to the factors found in the cyber deterrence material to answer the research question. The key observation presented in this study is that a comprehensive approach to building resilience in the society is essential for the credibility of cyber deterrence. Resilience in cyberspace should be viewed from the perspective of every layer, including logical, physical and cyber-persona layers.

Keywords: Cyber deterrence, Finland, Concept of Comprehensive Security, resilience

1. Introduction

Numerous cyber threats target critical infrastructure of our digitalised society. The impact of the threat is widened by the fact that these critical systems have interdependencies, so disruptions in one system can affect several others. (Lehto & Neittaanmäki, 2022) By developing society's resilience, the ability to anticipate, withstand and recover from disruptions to critical infrastructure is improved.

Finland has a Concept of Comprehensive Security (CCS) for building and sustaining resilience in Finnish society. Preparation for disruptive situations is carried out with the operating principle of overall safety, where society´s vital functions are protected by the authorities, the business world, organisations, and citizens. The growing importance of cyber security has led to emphasising the importance of cyber resilience in the CCS. (The Security Committee, 2017)

This study investigates the possibilities to utilise the CCS as a tool for creating deterrence in cyberspace and aims to create a new perspective on the international academic discussion of cyber deterrence. The type of resilience the CCS creates and its possibilities to be utilised in cyber deterrence are investigated. The CSS, as a model from the perspective of cyberspace, is also examined. This study leaves out of scope the practical application of the concept and the assessment of its functionality.

The research question is: *"How can the Concept of Comprehensive Security be utilised in building cyber deterrence?"* This question is answered by analysing the CSS material and theories of resilience and cyber deterrence.

Resilience as part of deterrence is a controversial topic. There are views according to which resilience and defence widen the scope beyond deterrence (Borghard and Lonergan, 2021). On the other hand, it can be thought that resilience can be used to build a basis for deterrence, because it signals to the potential aggressor that attacks won´t have the desired effect (Sweijs and Zilizny, 2020; Reire, 2016).

This study examines resilience as a part of deterrence from the perspective of cyberspace. Cyberspace is dealt with through a threefold division, where it is divided into physical, logical and personal layers (Laari et al., 2019). All of these layers must be considered when planning resilience and deterrence. Therefore, the protection of cyberspace cannot consist solely of technical means applied to the logical layer.

2. Concept for Comprehensive Security and its connection to cyberspace

The Finnish model, the Concept for Comprehensive Security (CCS), involves all society's actors to prepare for disruptive situations. The concept's role is more coordinating than managing, it creates conditions for cooperation for the authorities, businesses as well as citizens and non-government organisations. The concept is implemented in practice in branch-specific or cross-administrative strategies and executive programmes. (The Security Committee, 2017) The Finnish national cyber security strategy is closely linked to the CCS, and it relies on the general incident management model. (Turvallisuuskomitea, 2019)

Comprehensive security can be understood as *"a condition where threats and risks to the vital functions of society are prepared for"* (Sanastokeskus, 2017). This includes *"preparing for threats, managing disturbances and exceptional conditions and recovering from them"* (Sanastokeskus, 2017).

A vital function can be described as *"a function that is necessary for the functioning of society"* (Sanastokeskus, 2017). Critical infrastructure can be understood as *"basic structures, services and related functions that are necessary to maintain the vital functions of society"* (Sanastokeskus, 2017). The critical infrastructure consists of both physical structures and electronic functions and services. For example, energy production, transmission and distribution systems, transport and logistics, information and communication systems and water and waste management are all part of critical infrastructure. (Valtioneuvosto, 2022)

The CCS identifies seven vital functions in society that are to be protected: 1) leadership and management, 2) international and European Union cooperation, 3) defence capability, 4) internal security, 5) economy, infrastructure and security of supply, 6) functional capacity of the population and services and 7) mental crisis resistance. (The Security Committee, 2017)

Public administration services enable the leadership and the management of the state as well as the international and European Union cooperation. This includes the decision-making ability and shared situational awareness. These services include the state's common information and communication services, the security network and support services. ICT (Information and Communication Technology) functions and the production of digital services are national and international, enabling international cooperation. (The Security Committee, 2017)

Defence capability includes military defence, society's preparedness and national authority cooperation and international defence cooperation. The ability to receive and give military aid is part of Finland's defence capability. (The Security Committee, 2017) Military capabilities are dependent on information processing, the ability to form situational awareness and control of weapon systems and logistic systems. (Valtioneuvosto, 2021)

By maintaining the internal security, crimes against the population are prevented and combated, also consequences of accidents and environmental damage or other similar disturbances and threats are managed. Internal security includes border security, maintenance of public order and security, rescue operations and protection of the population in accidents and dangerous situations. (The Security Committee, 2017) Authority networks and communication and situational picture services, including emergency response systems, are critical in maintaining internal security. (Sisäministeriö, 2019)

The economy, infrastructure and security of supply consist of services and networks, such as financial transactions, information and communication systems, digital services and information, logistics services and electricity systems. These vital functions have many interdependencies in cyberspace. For example, financial market services are completely dependent on the functionality of telecommunication connections, ICT systems and the related electricity supply. (The Security Committee, 2017)

The functional capacity of the population and services includes systems of social work and health services, education and research systems. For example, in terms of social and health care, it is essential to secure diagnostic capability and the storage and transfer of patient and customer data. (The Security Committee, 2017)

Mental crisis resistance is the ability of an individual, community and society to withstand the mental pressure caused by crises and cope with their effects. Maintaining citizens' trust is essentially dependent on the communication of authorities, organisations and other communities in exceptional circumstances. The media plays a significant role in the nation's mental crisis resistance as the maintainer and creator of crisis resilience. (The Security Committee, 2017)

Table 1 summarises the vital functions and their connection to cyberspace mentioned in the CCS. It should be noted that these functions are not separate from each other, but they form an interconnected system where a malfunction in one could cause disturbances in one or several other functions (Valtioneuvosto, 2022).

Table 1: Vital functions in Finnish society and their connection to cyberspace.

Vital functions	Examples of connection to cyberspace
Leadership and management	Situational awareness, crisis communication, information sharing
International and European Union cooperation	Information systems and the Internet
Defence capability	Systems, networks and connections of the defence forces and authorities
Internal security	Authorities' systems, networks and connections, crisis and emergency communication, logistics, social and healthcare systems
Economy, infrastructure and security of supply	Financial transactions, information and communication systems, digital services and information, logistics, electricity
Functional capacity of the population and services	Systems of social work and health services, education and research systems
Mental crisis resistance	Crisis communication services, media

According to the Cyber Security Strategy 2013 (CSS13), Finland's vision of cyber security is the following: *"Finland can protect its vital functions in all situations against the cyber threat*" (Turvallisuuskomitea, 2013). These functions mentioned in the CSS13 are similar to the functions presented in table 1. The CSS13 states that national cyber resilience is measured in such a way that it can create preparedness by the goals of overall safety, the ability to function in cyber-disruption situations and recovery after cyber-disruptions and resilience. The goal is achieved by maintaining and developing the ability of companies and organisations to detect and combat cyber threats and disruptions, as well as to recover from them. Necessary protective capacity is maintained and resilience is developed for example by planning and using backup methods. (Turvallisuuskomitea, 2013)

3. Methods and results

This study investigates the resilience created by the CCS and how this resilience can be utilised in cyber deterrence. Content analysis (Puusa, 2021) was used as a research method. The material concerning resilience and cyber deterrence was gathered from the abstract and citation database Scopus. The search phrases included the terms *"cyber deterrence"* and *"resilience in cyberspace".* Purely technical articles were excluded, because they were out of the strategic perspective of this study. The search generated seventeen articles on cyber deterrence and nine articles on resilience in cyberspace. The material concerning the CCS was searched from the websites of the Finnish Government. This material consists of seven Finnish strategic documents. Analysed articles and documents are cited in the text.

Two hypotheses for the study were set based on previous cyber deterrence studies. Because deterrence implemented exclusively in cyberspace through cyber means has proven to be a difficult concept (Wei, 2015; Brantly, 2018), this study sets the first hypothesis as follows: *"deterrence in cyberspace is created by a wide selection of means".* In addition to cyber-specific means, these include political, economic, military and legal means, as well as resilience, which are used for defensive or offensive purposes (Brantly, 2018; Bendiek and Metzger, 2015; Burton, 2018; Chen, 2017).

The second hypothesis is that *"resilience is essential for cyber deterrence".* This was based on two assumptions: 1) not all cyberattacks can be prevented (Lindsay, 2015; Klimburg, 2012; Nye, 2017) and 2) the ability to recover can be used to demonstrate that the attack does not achieve the desired effects (Nye, 2017; Vogel et al., 2021; Kerttunen, 2019).

The first round of analysis consisted of the collected cyber deterrence material. The aim was to form an understanding of functional ways to create deterrence in cyberspace, supporting the first hypothesis. This view was deepened by studying the role of resilience in cyberspace in the second analysis round. The findings from the first two analysis rounds were compared to the third analysis round, where the CSS material was investigated from the perspective of cyberspace and its three layers. Gathering and comparing the results of these rounds formed the answer to the research question: *"How can the Concept of Comprehensive Security be utilised in building cyber deterrence?"*

3.1 Cyber deterrence

According to the hypotheses, cyber deterrence is best created with a wide selection of means that should not be limited just to cyberspace, although technical and cyberspace-targeted means should not be forgotten. These

means could include, for example, developing a regulatory framework against cyber-attacks and disruptions (Chen, 2017), developing resilience to recover from cyber-attacks (Vogel et al., 2021; Kerttunen, 2019; Burton, 2018; Tolga, 2018; Nye, 2017), creating international rules on acceptable activity and the legitimacy of targets in cyberspace (Chen, 2017; Bendiek and Metzger, 2015; Russell, 2015) and a national strategy for creating cyber deterrence (Bendiek and Metzger, 2015).

Resilience and functional security practices are likely to increase the cost of aggressions, therefore creating additional denial thresholds (Vogel et al., 2021; Kerttunen, 2019; Nye, 2017). This can be achieved, for example, through active cyber defence (Hurley and Watkins, 2016; Bendiek and Metzger, 2015). This requires real-time monitoring and threat detection capabilities, analysis capabilities and the ability to reduce the impact of threats. (Jasper, 2015)

Active cyber defence creates opportunities to deny the impact of the opponent's actions through resilience and to cause losses to the opponent by employing countermeasures. (Borghard and Lonergan, 2021; Jasper, 2015) Countermeasures should consider the possibility of kinetic retaliation if a cyberattack seriously damages the physical infrastructure or threatens the lives of citizens. (Chen, 2017; Bendiek and Metzger, 2015) Other means for retaliation are, for example, legal actions, financial sanctions or diplomatic isolation. (Burton, 2018)

Cyber deterrence needs to be created through the cooperation of different actors. These include both national and international organisations and authorities, as well as public and private sector actors. The deterrence created in this way is directed against state actors, threats within the state, as well as crime and terrorism. (Burton, 2018)

Both society and policymakers need to accept that there are areas where cyberattacks can´t be deterred. It should be noted that cyber deterrence needs prioritising, since cyber defence cannot be strong everywhere. (Kerttunen, 2019; Tolga, 2018; Lindsay, 2015; Klimburg, 2012) At a minimum, such cyber-attacks should be countered that have strategic effects by damaging national security and vital functions (Kostyuk, 2019). Cyber deterrence should not be treated as a separate strategy, but as a part of the state´s comprehensive deterrence and other strategies aiming to protect cyberspace. (Brantly, 2018; Lonsdale, 2018) For this, it is worth using existing structures and operating models. (Tolga, 2018)

In this article, cyber deterrence refers to the actions taken to convince an aggressor that an attack in cyberspace is not profitable, and it may cause counter-measures. The selection of means includes political, economic, military and legal means, as well as resilience. Therefore, cyber deterrence is created by the whole society, including authorities, as well as organisations and citizens. Cyber deterrence is understood to be part of the state´s comprehensive deterrence.

3.2 Resilience in cyberspace

Resilience can be understood as an internal characteristic of the object and as an interactive recovery process activated in times of crisis. (Hyvönen et al., 2019) Activities, such as anticipating, resisting, recovering, detecting and adapting to threats of any kind, are associated with resilience. (Vogel et al., 2021)

Resilience can be described as the ability of individuals and communities to *"maintain the ability to function in changing conditions and the readiness to face disruptions and crises and recover from them"* (Sanastokeskus, 2017). Resilience is used to respond to situations where safety is compromised by unexpected activity. (Sanastokeskus, 2017)

From a broader perspective, resilience can also be found in the attitudes and images of individuals, it can be seen in public debate and the shared values and goals of society. Resilience can be thought of as part of social capital. Making it public creates the citizen's trust in society and manifests itself, for example, in national narratives. (Rodin, 2015)

The resilience of information systems can be understood as *"the ability of an information system to operate under adverse conditions or stress, even if in a degraded or debilitated state, while maintaining essential operational capabilities, and to recover to an effective operational posture in a time frame consistent with mission needs*" (NIST, 2020). This includes two kinds of continuous capabilities, the first being the ability to continue services under disruption and the second the ability to restore the effective operating state within a tolerable time frame. (Zuo et al., 2021)

Most of modern society's functions are dependent on cyberspace. Cyber disruption situations can cause harm to several critical services and functions. Therefore, cyber resilience should be considered as a whole, which

includes physical, informational, cognitive and social dimensions. (Bellini and Marrone, 2020) Cyber resilience can be described as *"the ability to anticipate, withstand, recover from, and adapt to adverse conditions, stresses, attacks, or compromises on systems and it can still maintain a certain functional service capability*" (Zuo et al., 2021).

Cyber resilience can be approached as a combination of cyber security and society's crisis resilience. The continuity of society's critical services is ensured by the cooperation of several actors. These are governmental actors and those operating from critical infrastructure and online services. (Devanni et al., 2022; Burton, 2018)

Based on the definitions presented, resilience in cyberspace consists of actions that aim to preserve the system's vital functions or services in the event of a disruption and to restore or replace them within a desired timeframe. Resilience includes the ability to anticipate and detect disruptions and to learn from previous disruptions. It should also be noted that cyber resilience is not only technical or physical, because disruptions in cyberspace affect services used by people. The human dimension of cyber resilience includes mental resilience. In today's world, where information and services are immediately available, disruptions may cause mental pressure on the individual.

3.3 The Concept of Comprehensive Security and cyberspace

Understanding cyberspace is facilitated by dividing it into layers, each of which has a different role in processing information. This study adopts the Finnish view, which originates from the U.S. publication *JP 3-12, Cyberspace Operations* (US Joint Chiefs of Staff, 2018). This view divides cyberspace into three layers: physical, logical and cyber-persona.

The physical layer contains the geographical and physical network components used to store, transfer and process information. It includes hardware, systems and infrastructure that make up physical routes and networks. (Laari et al., 2019) From the critical infrastructure point of view, the physical layer includes the facilities where the equipment used by vital services, and the personnel who operate them, are geographically located. This also includes the physical devices involved in data transfer and data storage, for example, servers and sea cables. (The Security Committee, 2017)

The CSS builds resilience into physical cyber-infrastructure in various ways. Usability and safety of the necessary facilities and equipment are ensured in all situations. The supply of electricity is secured for critical services. Procurement chains are secured by national self-sufficiency, production and international agreements. Physical data transmission connections are secured and provisions are made for switching to backup systems. (Valtioneuvosto, 2022) The domestic cyber security industry is being developed to secure services. (Paananen, 2021)

The logical layer comprises logically programmable parts of the network. Logical components are not tied to a specific physical location. The layer consists of network settings, information security processes, data transfer protocols, internet domain names, ownership information and information, applications and protocols that control interaction with the physical layer. (Laari et al., 2019) From the point of view of critical infrastructure, the logical layer includes all the networks, systems and digital services which are needed for the vital functions of society. (Valtioneuvosto, 2022)

As a part of cyber resilience, the built-in security of national network services is being developed. The operation and safety of information reserves, services and systems critical to society are ensured throughout their entire life cycle. (Paananen, 2021) The interdependencies of the digital operating environment require an overall architecture that takes cyber security into account. (Security Committee, 2019)

The goal of cyber security is to identify and detect possible disruptions to vital functions and minimise the harmful effects of disruptions. (Security Committee, 2019) This also requires comprehensive cyber security situational awareness. (Paananen, 2021) Key players are building their resilience, including planning and practicing backup methods to operate under and recover from cyberattacks. Legislation must support the protection of society's vital functions and state security against cyber threats. (Turvallisuuskomitea, 2013)

The cyber-persona layer contains the virtual identities of individuals or communities in cyberspace. An individual or community can have simultaneously several cyber personalities. On the other hand, several people can share one cyber persona, for example, using a single e-mail account for many users. (Laari et al., 2019) From the perspective of critical infrastructure, the cyber-persona layer consists of virtual identities of personnel and organisations connected to the vital functions of society. Using these identities requires the person to have

sufficient information security knowledge and to be aware of possible threats targeted to usernames and passwords, for example, various means of social engineering. Also, people themselves are a critical part of the cyber-persona layer. (Laari et al., 2019)

The CSS builds resilience into the cyber-persona layer by developing the competence of individuals and organisations. (Turvallisuuskomitea, 2019) The goal is that every citizen, from a child to a pensioner, has sufficient skills to function in a digital society. Also, secure services must be nationally produced. Cybersecurity skills are needed both by the producers of the services and by the citizens using them. (Paananen, 2021)

Mental resilience is needed for the resilience of the cyber-persona layer, since severe cyberattacks could cause mental pressure on individuals. At the core of mental crisis resilience is the sharing of information through reliable communication and training. Organisations produce information for authorities and decision-makers about citizens' security needs, so communication works in both directions in an ideal situation. (The Security Committee, 2017)

The resilience to cyberspace produced by CCS is a complex entity that consists of the measures of many actors in society, security of supply, technology, backup systems and people's competence. Table 2 describes the results of the analysis described above.

Table 2: The CCS from the perspective of cyberspace.

Cyberspace layer	Critical infrastructure and vital functions connected to cyberspace	Resilience produced by the CSS
Logical	Networks, systems and digital services, which are needed for the vital functions of society.	Ability to detect, identify, analyse and minimise threats, situational awareness, backup methods, built-in security.
Physical	Facilities where the equipment used by vital services and the personnel who operate them are located. This also includes the physical devices involved in data transfer and data storage.	Usability and safety of facilities and equipment, secured electricity supply, national self-sufficiency, production and international agreements on material supply, secured data connections, backup systems, domestic industry.
Cyber-persona	People, cyber-identities of individuals and organisations.	Competence, mental resilience.

Figure 1 shows the impact of cyber resilience on cyber deterrence from the perspective of the three layers of cyberspace. Cyberspace is an integral part of society's critical infrastructure and functions, as they are connected to it in several ways. Resilience must be built into every layer of cyberspace, because a disruption in one layer affects the others.

The CCS considers the physical, digital and human aspects of society's critical infrastructure. Therefore, it creates resilience also for cyberspace that can be utilised for cyber deterrence. It should be noted, however, that from the cyber deterrence point of view, it is necessary to identify the aspects that can be used from the already existing cyber resilience, but also the gaps that the existing models do not respond to.

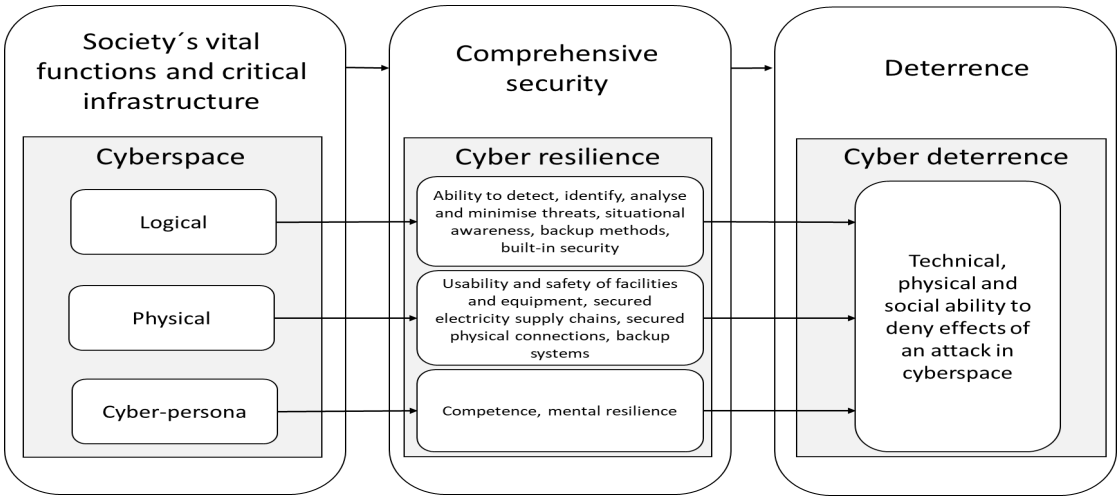


Figure 1: The CCS as a tool for cyber deterrence.

4. Conclusion

Looking at cyberspace as a whole, which includes a physical, logical and personal layer, resilience must be built into each of these. Most cyber deterrence studies focus on the logical layer and technical solutions, but the other two layers should not be forgotten. The resilience of the cyber-persona layer consists of the individuals and their cyber security expertise. The security of the physical layer is connected to the sovereignty of the state in the traditional sense.

Cyberspace affects society´s critical functions directly, because they connect to it in one way or another. By developing technical, physical and mental resilience, cyber disruptions are better tolerated. In Finland, CCS is an existing structure that can be publicly signalled to convince the aggressor of society's ability to withstand hostile actions. This creates deterrence, because the execution of the attack as planned or the planned impact is disputed.

Cyber deterrence is best created using all means available, including political, economic, legal and military capabilities. Because the adversary can always question the defender´s will to retaliate or attack, signalling resilience is a more plausible way to convince the opponent that the attack will not achieve the desired effects. However, for the sake of credibility and to be proactive, the state must also have the ability and publicly signalled will to retaliate.

It should be noted, that shortcomings have occurred in the practical application of the CCS, for example, during the COVID-19 pandemic (Tulevaisuusvaliokunta, 2020). Therefore, the results of this study must be viewed critically from the point of view that the functionality of the concept was not evaluated in practice. The subject of further research is whether the tasks and instructions given by the strategic documents can really be implemented according to the CCS.

The results of this study could be deepened in further research by investigating how narratives existing in society could be utilised in deterrence signalling, such as the importance of the resilience produced by CSS to society's crisis resistance. It is advantageous for the state to use functions of society already in place in deterrence signalling, even if they produce deterrence as a by-product.

References

Bellini, Emanuele and Stefano Marrone (2020). Towards a novel conceptualization of Cyber Resilience, *2020 IEEE World Congress on Services (SERVICES),* Beijing, China. DOI: 10.1109/SERVICES48979.2020.00048.

Bendiek, A., & Metzger, T. (2015). *Deterrence theory in the cyber-century*. Lessons from a state-of the-art literature review. https://doi.org/10.1007/s13347-017-0290-2

Borghard, Erica & Shawn Lonergan (2021). Deterrence by denial in cyberspace, *Journal of Strategic Studies*. DOI: 10.1080/01402390.2021.1944856.

Brantly, Aaron (2018). The Cyber Deterrence Problem. T. Minárik, R. Jakschis, L. Lindström (Eds.). *2018 10th International Conference on Cyber Conflict*. DOI:10.23919/CYCON.2018.8405009.

Burton, J., (2018). *Cyber Deterrence: A Comprehensive Approach?* Nato Cooperative Cyber Defence Centre of Excellence, Estonia 2018. https://ccdcoe.org/uploads/2018/10/BURTON_Cyber_Deterrence_paper_April2018.pdf

Chen, Jim (2017). Deterrence and its implementation in cyber warfare. In *International Conference on Cyber Warfare and Security*. Academic Conferences International Limited. https://www.jstor.org/stable/26633152

Devanny, Joe, Luiz Goldoni and Breno Medeiros (2022). Strategy in an Uncertain Domain: Threat and Response in Cyberspace. *Journal of Strategic Security 15*, no. 2. https://doi.org/10.5038/1944-0472.15.2.1954

Hurley, John and Lanier Watkins (2016). Cyberspace: The new Battlefield. In *11th International Conference on Cyber Warfare and Security*. Academic Conferences International Limited.

Hyvönen, Ari-Elmeri; Tapio Juntunen, Harri Mikkola, Juha Käpylä, Harri Gustafsberg, Markku Nyman, Tiina Rättilä, Sirpa Virta, Johanna Liljeroos (2019). *Kokonaisresilienssi ja turvallisuus: tasot, prosessit ja arviointi.* Valtioneuvoston selvitys- ja tutkimustoiminnan julkaisusarja 17/2019. https://www.fiia.fi/wp-content/uploads/2019/02/17-2019-kokonaisresilienssi-ja-turvallisuus.pdf

Jasper, S. (2015). Deterring Malicious Behavior in Cyberspace. *Strategic Studies Quarterly*, 9(1), 60–85. https://www.jstor.org/stable/26270834

Kerttunen, Mika (2019). Beyond Punishment: Deterrence in the Digital Realm. *Connections*, 18(1/2), 61–68. https://www.jstor.org/stable/26948849

Klimburg, Alexander (Ed.) (2012). *National Cyber Security Framework Manual*, NATO CCD COE Publication, Tallinn. https://ccdcoe.org/uploads/2018/10/NCSFM_0.pdf

Kostyuk, Nadiya (2021). Deterrence in the Cyber Realm: Public versus Private Cyber Capacity, *International Studies Quarterly,* Volume 65, Issue 4, December 2021. https://doi.org/10.1093/isq/sqab039

Laari, T., Flyktman, J., Härmä, K., Timonen, J. & Tuovinen, J. (2019). *#kyberpuolustus: Kyberkäsikirja Puolustusvoimien henkilöstölle*. National Defence University, Finland. https://urn.fi/URN:ISBN:978-951-25-3120-2

Lehto, Martti & Pekka Neittaanmäki (2022). *Cyber Security: Critical Infrastructure Protection.* Cham, Switzerland: Springer (Computational Methods in Applied Sciences). https://doi.org/10.1007/978-3-030-91293-2

Lindsay, Jon (2015). Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack, *Journal of Cybersecurity*, Volume 1, Issue 1, September 2015. https://doi.org/10.1093/cybsec/tyv003

Lonsdale, David (2018). Warfighting for Cyber Deterrence: A Strategic and Moral Imperative. *Philosophy & Technology*. 31. DOI: 10.1007/s13347-017-0252-8

National Institute of Standards and Technology NIST (2020). *Security and Privacy Controls for Information Systems and Organizations.* Special Publication 800-53, Revision 5. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf

Nye, J. (2017). Deterrence and Dissuasion in Cyberspace. *International Security 2017*; 41 (3): 44–71. DOI:10.1162/ISEC_a_00266

Paananen, Lauri (2021). *Kyberturvallisuuden kehittämisohjelma.* Liikenne- ja viestintäministeriö, Helsinki Finland. ISSN 1795-4045 pdf. http://urn.fi/URN:ISBN:978-952-243-599-6

Puusa, A. & Juuti, P., (2020). *Laadullisen tutkimuksen näkökulmat ja menetelmät*. Helsinki: Gaudeamus.

Reire, Gunda (2016). Resilience Challenges in the Baltic Countries. In Andžāns M, Bruģe I (eds) *The Baltic Sea Region: Hard and Soft Security Reconsidered*. Latvian Institute of International Affairs, Riga.

Russell, Alison (2015). Strategic Anti-Access/Area Denial in Cyberspace. *2015 7th International Conference on Cyber Conflict.* NATO CCD COE Publications, Tallinn. https://ccdcoe.org/uploads/2018/10/Art-11-Strategic-Anti-Access-Area-Denial-in-Cyberspace.pdf

Rodin Judith (2015). *The Resilience Dividend; Managing Disruption, Avoiding Disaster, and Growing Stronger in an Unpredictable World*. Profile Books, London.

Sanastokeskus (2017). *Vocabulary of Comprehensive Security*. https://turvallisuuskomitea.fi/wp-content/uploads/2018/02/Kokonaisturvallisuuden_sanasto.pdf

Sisäministeriö (2019). *Kansallinen riskiarvio 2018*. Sisäministeriön julkaisuja 2019:5. https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161332/5_2019_Kansallinen%20riskiarvio.pdf

The Security Committee (2017). *The Security Strategy for Society*, Government Resolution / 2.11.2017, Finland. https://turvallisuuskomitea.fi/wp-content/uploads/2018/04/YTS_2017_english.pdf

Tolga, Ihsan Burak (2018). *Principles of Cyber Deterrence and the Challenges in Developing a Credible Cyber Deterrence Posture*, NATO Cooperative Cyber Defence Centre of Excellence, Tallinn. https://ccdcoe.org/uploads/2018/10/Challenges_in_Developing_Credible_Cyber_Deterrence_Posture_in_Cyberspace-1.pdf

Turvallisuuskomitea (2013). *Suomen kyberturvallisuusstrategia.* Finland. https://turvallisuuskomitea.fi/wp-content/uploads/2018/05/Suomen-kyberturvallisuusstrategia-ja-taustamuistio.pdf

Turvallisuuskomitea (2019). *Suomen kyberturvallisuusstrategia 2019.* Finland. https://turvallisuuskomitea.fi/wp-content/uploads/2019/10/Kyberturvallisuusstrategia_A4_SUOMI_WEB_300919.pdf

Tulevaisuusvaliokunta (2020). *Koronapandemian hyvät ja huonot seuraukset pitkällä aikavälillä.* Eduskunnan tulevaisuusvaliokunnan julkaisu 1/2020. https://www.eduskunta.fi/FI/naineduskuntatoimii/julkaisut/Documents/tuvj_1+2020.pdf

Sweijs, Tim & Samuel Zilincik (2020). The Essence of Cross-Domain Deterrence. In Osinga, Frans & Tim Sweijis (eds). *Deterrence in the 21st Century—Insights from Theory and Practice*. Netherlands Annual Review of Military Studies 2020. https://doi.org/10.1007/978-94-6265-419-8_8

US Joint Chiefs of Staff (2018). *JP3-12 Cyberspace Operations*. https://irp.fas.org/doddir/dod/jp3_12.pdf

Valtioneuvosto (2022). *Valtioneuvoston huoltovarmuusselonteko*. Työ- ja elinkeinoministeriö. http://urn.fi/URN:ISBN:978-952-383-803-1

Valtioneuvosto (2021). *Valtioneuvoston puolustusselonteko*. Valtioneuvoston julkaisuja 2021:78. http://urn.fi/URN:ISBN:978-952-383-820-8

Vogel, Elisabeth, Zoya Dyka, Dan Klann, and Peter Langendörfer (2021). "Resilience in the Cyberworld: Definitions, Features and Models" Future Internet 13, no. 11: 293. DOI:10.3390/fi13110293

Wei, Lee Hsiang (2015). The Challenges of Cyber Deterrence. *Pointer, The Journal of the Singapore Armed Forces*. Vol.41 NO.1. ISSN 2017-3956. https://www.mindef.gov.sg/oms/safti/pointer/documents/pdf/POINTER_Vol41_1.pdf

Zuo, Jinxin, Ziyo Guo, Jiefu Gan and Yueming Lu (2021). Enhancing Continuous Service of Information Systems Based on Cyber Resilience, *2021 IEEE Sixth International Conference on Data Science in Cyberspace (DSC)*, Shenzhen, China. DOI: 10.1109/DSC53577.2021.00085.

Building Situational Awareness of GDPR

Pauliina Hirvonen and Martti J. Kari
University of Jyväskylä, Finland

pauliina.a.hirvonen@student.jyu.fi
martti.j.kari@jyu.fi

Abstract: Because previous academic research does not comment sufficiently on how the relevant content of the European Union (EU) General Data Protection Regulation (GDPR has been properly communicated to the organisations, or how the situational awareness (SA) of GDPR has been built in the organisations, this qualitative empirical research was regarded as a valuable approach for gathering authentic research material on the practical bases of this phenomena. The aim of this empirical case study (CS) is to develop a picture of what processes organisations use to build SA of the GDPR requirements. To guide the CS, we asked how the SA for decision-making was constructed and how it was perceived in organisations. The experiences of eight Finnish organisations showed that the organisations' practices of building SA and their experiences with the quality and adequacy of SA differed. However, building SA proved to be a critical step for organisations in the overall process of meeting GDPR requirements. Especially the data coming from inside the organisation became very relevant in the SA process, because it supported decision makers to determine how the GDPR requirements should be implemented in the organisation. As a main contribution of this article, based on best practices shared by organisations a model of building SA was built. The proposed model is threefold and was constructed by combining the findings of an empirical CS analysis, the steps of the intelligence process, and the essential elements of the model of creating information security SA. The result is potentially beneficial for building situational understanding of any complex or ambiguous issue, especially in complex and digitalised technological areas, where combining information management with accurate and efficient decision-making is a common challenge. The results can be used by any party who is looking to build SA of an abstract issue in a complex environment.

Keywords: GDPR requirements, Interpreting privacy regulation, Understanding GDPR requirements, Organisational situational awareness, Privacy practices, Information security development

1. Introduction

Previous research does not comment sufficiently on how the relevant content of the European Union (EU) General Data Protection Regulation (GDPR) has been properly communicated to the organisations, or how the situational awareness (SA) of GDPR has been built in the organisations. As, for example, Ahmad et al. (2021) stated, most research has focused more on technological than practical perspectives. The development history of the research topic has been significantly influenced by the need to produce practical research information on GDPR issues in academic research, which can support the environment, for example service providers or legislators to develop usable solutions for organisations that facilitate the implementation of privacy, for example GDPR issues. GDPR was perceived by organisations as impractical and challenging from the start (see e.g. Tankard, 2016; Sirur et al, 2018; Hirvonen, 2022a and 2022b), so clarifying solutions and a model are needed to improve developing GDPR SA. The research data for this case study was already collected in 2021, so more scientific research on the topic has naturally been published after that. However, it can be stated that even after 2021, research progress does not concentrate only on the privacy side, but more broadly on e.g. cyber security, although even these solutions can be seen to benefit the privacy side as well.

The aim of this case study (CS) is to develop an overall picture of how organisations build SA of GDPR requirements. The objectives are 1) to collect practices found to be working by organisations, and 2) based on these to provide a model with which other organisations can facilitate when creating the SA. The research question is: How was the SA for decision-making constructed and how was it perceived within the organisation?

The concept of SA has several definitions. According to Tikanmäki and Ruoslahti (2019), all SA definitions share the idea of knowing and understanding what is occurring, prediction of the future situation, and the decision-making based on these factors. In this study, SA means the organisational awareness of all the actions, decisions, expenditures, resources, investments and mindsets required to demonstrate GDPR compliance in practice. SA also essentially involves an understanding of the changes in future needs of the organisation, the environment and GDPR. Organisations use their SA as the basis for decisions about the activities and resources related to GDPR.

The results here indicate that organisations' practices for building SA and their experiences of its quality and adequacy differed. However, building SA proved to be a critical step for organisations in the overall process of

meeting GDPR requirements, because decisions on the necessary actions and investments are made based on SA. Using these results, necessary parts for building a model of building optimal SA for organisational decision-making is gathered. This research of building SA of GDPR has the following structure: the highlights of existing literature are introduced in Section 2, the methodological choices are shortly described in Section 3. The results and analysis are presented in Section 4. The discussion is in Section 5. Finally, the conclusions are located in Section 6.

2. Early Work

Quite a limited amount of research of what is already known about building the SA, concentrating specifically on the GDPR or data protection (DP) requirements approach, was found. For example, Teixeira et al. (2019) acknowledged the importance of regulation awareness in organisations, including requirements and obligations. Their review, which examined the success factors of GDPR implementation, was the most complete research of the ones discussed here, because it highlighted research issues comprehensively, contributing the identification of enablers, benefits, challenges and the barriers affecting the compliance process. Aberkane et al. (2021) investigated the consideration of the GDPR regulation when developing service solutions in organisations. Organisations' challenges with interpreting GDPR have been reported by Tankard (2016) and Sirur et al. (2018). As also the empirical evidence demonstrated that in 2018 more than 50% of all websites failed to present privacy policies for consumers in an adequate manner (Dellinger, 2019). Seo et al. (2017) warned of forthcoming challenges due to the GDPR for organisations that process large amounts of personal data. Capgemini's survey (2018) also showed, in practice, that most of the companies in Europe and the United States were not ready in time for GDPR. Surveys by Dell (2016), KPMG (2017), Addis and Kutar (2018), and Presthus et al. (2018) noted the lack of SA of GDPR to be common in many organisations.

SA was examined widely from other aspects than GDPR. For example, cyber SA was reviewed by Franke and Brynielsson (2014) and Al Sagri and Al Aboodi (2015). E-health privacy risk awareness was surveyed by Bellekens et al. (2016). Pöyhönen et al. (2019) examined cyber SA in critical infrastructure organisations. Argaw et al. (2020) addressed the cybersecurity of hospitals, noting that information sharing simplifies SA and a clear perception of threats and details related to threat actors. Rajamäki and Katos (2020) examined information-sharing models related to cybersecurity early warning arrangements, highlighting the observation that operating protection services as well as for emergency and crisis management were dependent on common cyber SA. They (Rajamäki & Katos, 2019) also observed data analytics techniques (e.g., clustering and classification) promoting SA, when integrated as a part of the information sharing system. Rendall et al. (2021) used a survey to examine cyber threat detection SA. Panagiotis et al. (2019) noticed that Internet of Things (IoT) applications creation requires advanced knowledge extraction and real time situations. Endsley (2000) presented the Situation Awareness Global Assessment Technique (SAGAT) based on the SA requirements of the operator. Sanneman and Shah (2020), in turn, examined the application of SAGAT in measuring the situation awareness of end-users that use autonomous intelligent agents. SA has also been studied from the following perspectives: geo-distributed situation awareness applications in the fog by Saurez et al. (2016), SA related to port security by Adams et al. (2020) in the European Commission's project Scalable multidimensionAl sitUational awaReness sOlution for protectiNg european ports (SAURON), and an intrusion detection system for network security SA by Mahendiran and Appusamy (2018). Vieweg et al. (2014) highlighted that information broadcast through social media may clarify SA during a crisis.

Most recent cyber SA research has been conducted after the research data for this case study was collected in 2021. For example Rowan et al. (2021) leveraged SA approach to examine decision autonomy when individuals concerned the permission for eConsent and Renaud and Ophoff (2021) conducted a SA model to improve the implementation of Cyber security controls and precautions. Ahmad et al. (2021) conducted a model that guides organisations to build a SA of cyber threats and business context emphasising the incident response approach. Insident response side was examined also by Husák et al. (2022). Toxirjonovich and Tulanovna (2022) brought forth the gaps and opportunities related to cyber SA. Technological side approach was represented by Apostolakis et al. (2022) proposed a solution for SA based on Augmented reality (AR) and artificial intelligence (AI), and Chung et al. (2023), Jiang et al. (2002) and DeValk (2022), that examined the visualisation opportunities related to Cyber SA. Schaberreiter et al. (2022) proposed a cybersecurity SA solution for local public administrations referred to developing European legislator cybersecurity framework. They (Schaberreiter et al., 2022) performed the soft systems methodology to support collection, analysis and visualisation processes related to SA process. The comprehensive research (Schaberreiter et al., 2022) is funded by European commission and the research is a part of a project that aimed to create a set of software tools to support the

fight against the use of social media by terrorist organizations and describes for example the layered application architecture for the solution. So after these examples it can be concluded that the mount and types of research of cyber SA has increased in the past few years.

The review produced examples of perspectives on SA from the literature during the last decade. It can be concluded that SA research is diverse and includes both empirical and theoretical research from different fields. However, the previous research does not detail how precisely the SA of GDPR or data DP requirements has been built by the organisations.

3. Methodology

The CS method is used in this article. Bennett (2004) recognised the following aspects as the main advantages of using CS compared to other methods: 1) the notice of new variables and hypotheses and the variables that impact on causal mechanisms, 2) the discovery of history-based explanations for the cases, 3) the increase of the construct validity, and 4) the possibility to use generalisations to model complex relationships. The CS approach was chosen for this study precisely to enable the identification of factors influencing the cause-and-effect relationship related to SA process.

The representatives in charge of GDPR matters in eight either large or medium-sized Finnish organisations with varying industry were interviewed. Interviews were semi-structured thematic interviews and the following 14 open-ended questions presented in table 1 were discussed:

Table 1. **The Interview Questions.**

1. What kind of process was used to build SA of GDPR requirements in the organisation?
2. Who in the organisation builds or utilises SA of GDPR requirements?
3. From which sources outside your organisation have you received useful support for the interpretation or implementation of GDPR?
4. What procedures were used to interpret the theoretical GDPR requirements into practical actions and measures for the organisation?
5. On what basis can the reliability and accuracy of information and data sources related to GDPR requirements be assessed? What kind of information has been relevant and useful?
6. Describe the organisation's experience with GDPR guidance provided by official governing bodies during the process.
7. What kinds of deviations have occurred from the official guidelines during the process?
8. Describe the importance of organisational initiative and activism in the process of building SA of GDPR requirements.
9. How was GDPR-related data management handled in the organisation? What effects has GDPR had on organisational data management?
10. In your opinion, what are the possible causes of occasional inaccuracies or deficiencies in the organisation's SA of GDPR?
11. What has been the most challenging step in building or maintaining SA of GDPR?
12. Are there any differences between building SA of GDPR and other processes in the organisation?
13. What are the most important factors for an organisation to be able to build a correct, comprehensive and reliable SA of GDPR requirements?
14. Describe a successful GDPR-related organisational leadership culture.

4. Results and Analysis

Data was analysed using explanation-building mechanisms, cross-case comparisons, and themes. The kind of set was chosen based on the type of the research question and case types. The results showed that the SA processes for organisations' GDPR requirements varied. Eight types of processes were identified. Typically, the SA process involved becoming familiar with training GDPR content, gathering information within organisations, creating processes, preparing guidelines, training, evaluating, and developing processes and personnel. The processes shared a central focus on the continuous interpretation and integration of data from within the organisation with data from inside and outside of it. Surprisingly, the data coming from inside the organisation became very relevant in the SA process, determining *how* the requirement was implemented in the organisation. The external data, in turn, formed the frames of *what* to do. The core of organisations' efforts during the SA process targeted the organisation internally in implementation awareness and compliance in practice.

According to the responses, the GDPR SA content is typically produced within the organisation at four levels: 1) all staff (day-to-day consideration of GDPR in one's own activities), 2) line organisation level/supervisors (project ownership), 3) DP organisation (GDPR coordination) and 4) executive management (ultimate responsibility, permission and resources for larger projects). The study identified third parties from whom organisations had

received assistance during the GDPR process. It turned out that the understanding of GDPR requirements was influenced by the organisation's own understanding, consultant support, and clarity of authority. It was found that the procedures of organisations to interpret GDPR requirements as concrete actions varied. Assessing the reliability of GDPR requirement information is seen as important, especially for information coming from outside the organisation, as there exists misinformation, misinterpretations, and different levels of service providers (e.g., experienced experts). The information that was the most reliable, but which required interpretation, was considered to be the instructions of the official governing bodies. The attributes of usable control information were defined as follows: unambiguous, filtered, clear, and pragmatic listing that meets customer needs. SA building was also hampered by the following reasons: official guidance information required analysis and was challenging to interpret, not all requests of information are answered, and there are known delays and precedents. Trusted partners were perceived by three interviewees as relevant in the successful interpretation of the requirements. One interviewee also mentioned that more official and more practical control information only became available about 18 months after the end of the transition period (2019–2020), but there was still ambiguity in the guidelines at the end of 2021.

The understanding of what information is in different systems and how it moves was emphasised. The following attributes were referred to as good information management: centralised, in one place, formal, readiness for change, clear communication through different channels, understanding of needs, emphasis on access control, and a system for highlighting important issues (e.g., a ticketing system). The intermittent inaccuracies in SA were mainly due to a lack of understanding of the cross-section of the GDPR in the organisation, shortcomings in formal management communication, lack or inaccuracy in the interpretation of the information collected, and changes and new needs in one's own organisation. Several respondents saw interpreting and coordinating the organisational needs and requirements and business impact assessment (especially new functions) as the most challenging aspects of the SA process.

In the SA process, compared to other processes, the continuous activity and initiative of the organisation was emphasised, as was the fact that it is not a separate but a day-to-day and closely embedded issue for the whole organisation, which requires all staff to understand what is being done. The need for interpretation, ownership, cooperation, support and close business involvement were emphasised. The process of building SA cannot be totally outsourced because outside parties do not know enough about the organisation. The most significant enablers for the success of the SA-building process were identified as the mandate and practical resources provided by executive management (time, people, money), organisational knowledge, careful interpretation of "control data", accountability, commitment, professionalism in GDPR content, training and auditing. Successful leadership required process orientation, clarity (communication, roles, responsibilities, rights), the ability to link requirements to organisational operations and the identification of key personnel, contract management skills, activity, responsibility and overall commitment, as well as strong training for the executive team.

5. Discussion

The model of building optimal situational awareness is proposed here. This threefold model is constructed by combining the findings of an empirical CS analysis, the steps of the intelligence process, and the essential elements of the model of creating information security SA presented by Puhakainen (2016). Generally, the term "intelligence" is used mostly in military science, psychology and technology development and has several definitions. Here the term refers to a process that produces information about an object and circumstances that can then be used to build SA and support decision-making. Puhakainen (2016) examined more than 50 approaches to building information security awareness and, on the basis of his analysis, formulated theory-based IS security awareness (Isec SA strategy). The main elements of the strategy, namely the cognitive and behavioural approaches (Puhakainen, 2016) are used in this model to ensure efficient information security development while building SA of the GDPR requirements. The model is presented in Figure 1 containing the following phases:

1) Data Collection
2) 2-Step Analysis
3) Guidance and Control
4) Distribution
5) Evaluation and Quality

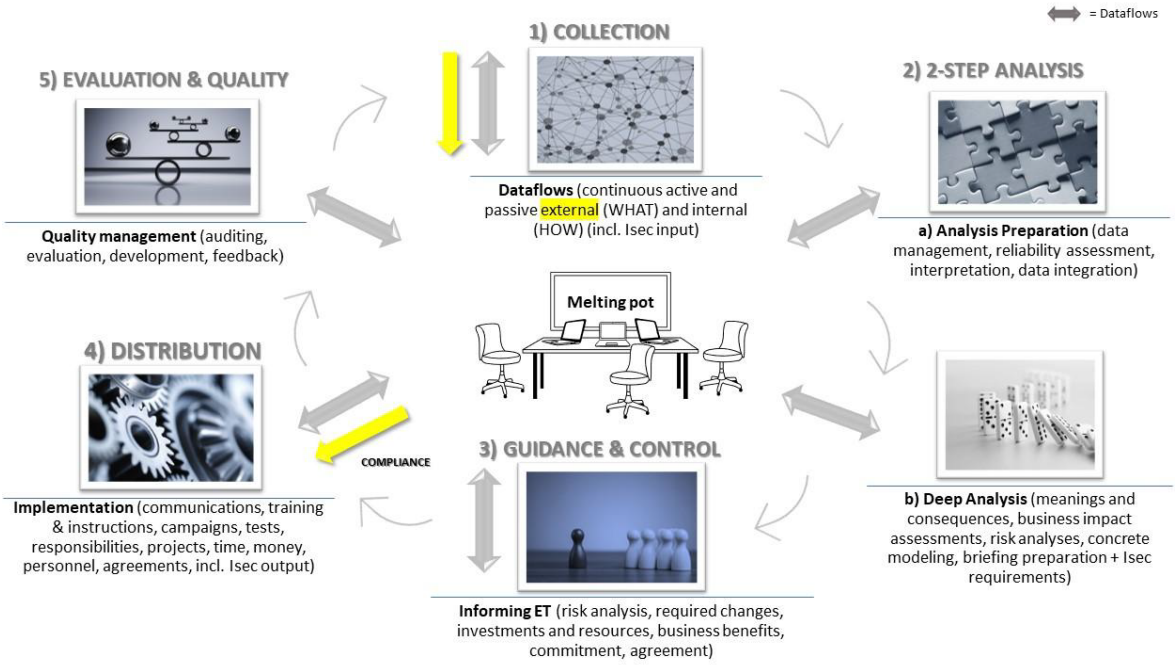


Figure 1. The Model for Building SA of GDPR Requirements.

At the heart of the SA model is the body responsible for building and maintaining SA (DP "Melting Pot"), which actively and continuously collects, processes and shares understanding, gives responsibilities and engages actors in two-way communication. In addition, executive management is kept informed and involved in supporting and enabling the process at all times. Information security demands are involved throughout the process. The following is a step-by-step overview of the model.

Table 2. Description of the Phases of Building SA.

1) Data Collection
Information flows contain information and data from various channels, with differing formats and content. These flows can be actively collected, where the organisation obtains the data through its own activities or, for example, by requesting them from others as a result of collection. These other parties can include industry umbrella organisations, legal partners and so on. Information flows come from outside the organisation (*what* to do) as well as from within *(how* things are done) and identifying the reliability of these flows and their sources is particularly important. External information flows include first-hand sources (EU and national DP authorities), law companies and experts, interbranch organisations, external auditors, ICT service providers, contractors, various training providers, media specialists in DP, external consultants, the partner network, advocacy organisations and peer-to-peer networks. The information flow from within an organisation should be two way: information comes in and enters a process in a fast cycle (e.g., needs, guidance, feedback, questions, requirements, risks). It is important to differentiate, on the one hand, between the layers within the organisation from which different information comes and, on the other hand, those who also need different levels and types of information for their use. Within the organisation, there are five levels, between which the information flows are key: 1) executive management, 2) DP organisation, 3) line management (process and function owners) and 4) all personnel, and the 5) organisation's Isec and Csec functions. If the information flow is not directed from the DP organisation to the organisation at the data collection phase, an information cycle conflict arises, the result of which is that the information flows within the organisation will be incomplete.
2) 2-Step Analysis
Preparatory analysis. At the beginning of the preparatory phase, the key is to ensure the reliability and usability of the data, which means assessing the reliability of the data content and structuring the different forms of data into a specific format to make them usable from both an analytical and a data management perspective. Errors can occur when integrating internal and external data. The data may be corrupt, dirty, old, or misdirected. A properly designed and frequently updated collection plan will partially address these challenges. Errors may also occur in the analysis. The individuals involved in data interpretation and integration are prone to cognitive skew and heuristic distortions. Cognitive skew is a psychological concept, which means that a person tends to perceive and emphasise perceptions, interpretations, and information in certain ways that lead to the wrong result. At worst, cognitive biases can prevent an analyst from perceiving the situation correctly even if all the data and information are in use. Cognitive biases include reinforcement bias, mirror image bias, and vitality bias. In confirmation bias, an individual advocates for information that supports their own preconceptions, gathers evidence, and remembers things selectively. In mirror image illusion, the individual estimates that everyone is acting in the same way as they are. When a situation is hectic, attention is drawn too early and too closely to the "busy" scenario. Heuristics are simplified choice models and shortcuts to thinking requiring as little cognitive effort as possible and are a quick and effortless way to make decisions. A good heuristic is a quick and near-optimal answer, but a bad heuristic, or heuristic distortion, violates logical principles and can lead to wrong conclusions. Heuristic distortions include anchor bias, association bias, and

availability bias. In anchor bias, the wrong initial situation or initial value is taken as the basis for further analysis. In association bias, the transaction is unjustifiably linked to weak or first-come, first-served evidence. In access bias and availability bias, the most readily available data and information are used. Cognitive skew and heuristic distortions in the SA process of GDPR requirements can be prevented and reduced by using the right analytical tools suitable for that analysis. The aim is to remedy these problems by requiring the analyst to be objective in all cases and by training the client to accept or not draw conclusions based on the data generated by the integration. Information must be interpreted in such a way that it is an unambiguous, filtered, clear and pragmatic listing that meets the needs of the user. This includes coordination and impact assessment of the organisation's needs and requirements for the business. It also encompasses Isec requirements as part of other activities, with the assistance of an Isec representative in practical issues (processes, system procurement, upgrades and modifications, training, development, policy, tools, etc.). At the end of the preparation phase, structurally uniform data streams from different sources will be integrated together, including Isec needs.
In-depth analysis. In-depth analysis involves the analysis of combined data, which means interpreting and processing the data into a holistic understanding and evaluating its meanings and impacts. At this stage of the analysis, the interpretation is also based on the assistance of experts. In-depth analysis includes, for example, assessments of the impact and significance of information, risk analyses, selection of priorities, preparation of justifications for development proposals and preparation of these findings for presentation to executive management. Concretised representations of Isec needs will be integrated into the package at this stage.
3) Guidance and control
The findings of the analysis are reported to executive management to make their commitment, support, authorisation and agreement, necessary refinements and practical resources (if needed) available for practical implementation. Also, a plan for integrating the GDPR systematically into business will be made. The necessary changes and additions will be made before final approval and implementation. The commitment of executive management is obtained through training and briefings. Errors related to the distribution of information may include, for example, the analyst modifying the result of the analysis to the liking of a supervisor or other client. Another challenge in distribution is that the customer does not understand or does not want to understand the results of data integration.
4) Distribution
This distribution occurs through communication, creation of guidelines, training, contracts with Isec service providers and other contractors, processes and necessary actions. It is necessary to determine what actions, needs, and responsibilities the implementation includes in practice, what each level of the organisation needs, and what is produced outside the organisation. Staff engagement is implemented through training, division of responsibilities, clear goals, and active two-way communication throughout the process.
5) Evaluation and Quality
Quality management includes, for example, evaluation, development, data management and quality management itself as integral parts of the whole SA process. Emphasis is placed on two-way and active information flows, and in particular on communication within the organisation, observation, feedback and questions, as well as on training and evaluation with the different practices and tools.

Compared to the traditional intelligence process, this model emphasises more open two-way communication as well as commitment and division of responsibilities between different actors, alongside step-by-step analysis before decision-making and guidance. In the case of an organisation's embedded function, the model also emphasises interpretation, ownership, support for collaboration, and the close involvement of the business. The focus of the model is on interpreting and coordinating requirements from within the organisation with respect to external GDPR requirements before making and guiding decisions. During the SA process, the focus inside of the organisation is on implementing practical awareness and compliance while collecting, analysing and reporting information and developing the process. and it is proposed to be leveraged in any abstract issue in a complex environment.

5.1 Implications for Practice

Puhakainen (2016) defined a concept for design theorising regarding Isec SA, which included the appropriate kernel theories, offering instructions for behavioural change, and establishing a testable research approach for scholars. This kind of approach is presented here regarding SA of GDPR requirements. Our article contributes a mixed epistemic–pragmatic model for building SA of any abstract matter in a complex environment. This model enhances the initiation of the necessary actions, reduces the delay in decision-making and embeds the awareness-raising in the organisation's operations holistically. This is realised by, for example, emphasising a more active and integrated relationship between the interpretation and operationalisation of information than what exists in the traditional organisational process. Step-by-step implementation instructions are also described for any organisations that need to build SA of any abstract issue. The model is also a suggested tool for DP, Isec - and Cyber security (Csec) managers that are integrating these security aspects into a single entity. The model is also testable for researchers in future studies. The result is potentially beneficial for building situational understanding of any ambiguous issue, especially in complex and digitalised technological areas, where combining information management with accurate and efficient decision-making is a common challenge.

5.2 Restrictions and Future Study

In this study, data was analysed by using comprising and theming. However, in the comparative study like this, the simultaneous use of the other methods to observe the research problem would have supported more easily to assess the accuracy of the chosen method. Rihoux (2006) analysed the development of the qualitative comparative analysis (QCA), noting that QCA and other related techniques are compatible especially with comparative historical analysis and theory-led case-oriented research. In the future, to support the evaluation of the accuracy of the method could for example these kinds of approaches be considered. One typical limitation of the CS approach relates to a lack of generalisability of the results. The research included experiences from only eight Finnish organisations and there may be differences in experiences in, for example, other EU countries. The larger the samples that are examined and analysed, the more valuable the approaches will be, and overall understanding of the issue will increase. In the future, to fill in the gaps left by limited amounts of empirical research, more empirical evidence of the factors that impact building SA of GDPR requirements is needed. Moreover, more experiences of different SA process building approaches are needed to develop an understanding of the best practices concerning the issue. The model presented here needs to be empirically tested and further experiences are needed to develop it.

6. Conclusions

Previous research has not fully addressed whether organisations have had sufficient SA of GDPR requirements. The aim of this CS was to clarify the landscape of how SA of GDPR requirements is built from an organisational perspective. To guide the CS, we asked how the SA for decision-making was constructed and how it was perceived in eight organisations. The experiences of eight Finnish organisations showed that the organisations' practices of building SA and their experiences with the quality and adequacy of SA differed. Overall, as many different processes were identified as there were respondent organisations. The essential part of the processes included the continuous interpretation and integration of data from within the organisation with data from inside and outside of it. The internal organisational data determined how the requirement was implemented in the organisation and the external data impacted of what to do. The SA process emphasised focusing on the developing organisational implementation awareness and compliance in practice. However, building SA proved to be a critical step for organisations in the overall process of meeting GDPR requirements. Therefore, the main contribution of this research was a mixed epistemic–pragmatic model of building SA. This model, with its detailed implementation instructions, could possibly be of benefit to those building SA of any abstract issue in a complex environment.

References

Aberkane, A. J., Poels, G. and Broucke, S. V. (2021). Exploring Automated GDPR-Compliance in Requirements Engineering: A Systematic Mapping Study. *IEEE Access*, (9), pp. 66542-66559.

Adams. N. P. H., Chisnall, R.J., Pickering, C. and Schauer, S. (2020). How port security has to evolve to address the cyber-physical security threat: lessons from the Sauron project. *Int. J. Transp. Dev. Integr.*, (4:1), pp. 29–41. WIT Press. https://doi.org/10.2495/TDI-V4-N1-29-41

Addis, M.C. and Kutar, M. (2018). The general data protection regulation (GDPR), emerging technologies and UK organisations: awareness, implementation and readiness. *UK Academy for Information Systems Conference*, pp. 1-23.

Ahmad, A., et al. (2021). How can organizations develop situation awareness for incident response: A case study of management practice. *Computers & Security*, 101, 102122.

Al Sagri, H.S. and Al Aboodi, S.S. (2015). Privacy awareness of online social networking in Saudi Arabia. *Cyber Situational Awareness, Data Analytics and Assessment* (CyberSA), *2015 International Conference*, pp. 1–6.

Apostolakis, K. C. et al. (2022). *DARLENE–Improving situational awareness of European law enforcement agents through a combination of augmented reality and artificial intelligence solutions.* Open Research Europe, 1, 87.

Argaw S.T., Troncoso-Pastoriza, J.R., Lacey, D. et al. (2020). Cybersecurity of Hospitals: discussing the challenges and working towards mitigating the risks. In *BMC Med Inform Decis Mak* (20), p. 146.

Bellekens, X., Hamilton, A., Seeam, P. et al. (2016). Pervasive ehealth services a security and privacy risk awareness survey. *2016 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (CyberSA).* IEEE, pp. 1-4.

Bennett, A. (2004). Case study methods: Design, use, and comparative advantages. *Models, numbers, and cases: Methods for studying international relations*, 2(1), pp. 19-55.

Capgemini. (2018). 85% of firms struggle to comply with GDPR by deadline, but opportunity exists for organizations who get it right. *Capgemini's Digital Transformation Institute´s survey: "Seizing the GDPR Advantage: From mandate to high-value opportunity"*. https://www.capgemini.com/no-no/news/85-of-firms-struggle-to-comply-with-gdpr-by-deadline-but-opportunity-exists-for-organizations-who-get-it-right/

Chung, M. H. M. et al. (2023). *Enhancing cybersecurity situation awareness through visualization: A USB data exfiltration case study*. Heliyon, e13025.

Dell. (2016). *GDPR: perceptions and readiness: a global survey of data privacy professionals at companies with European customers*. www.eurocloud.fr/wp-content/uploads/2016/10/gdpr.pdf

Dellinger, A.J. (2019). A Year Later, Many Sites Are Still Failing to Meet Basic GDPR Requirements. https://www.forbes.com/sites/ajdellinger/2019/05/31/a-year-later-many-sites-are-still-failing-to-meet-basic-gdpr-requirements/?sh=5025963d1eb9.

DeValk, K. S. T. (2022). *Real-Time Cybersecurity Situation Awareness through a User-Centered Network Security Visualization*. Doctoral dissertation, University of Maryland, College Park.

Endsley, M. Direct measurement of situation awareness: validity and use of SAGAT. In Endsley, M. R. and Garland D. J (Eds.) (2000) *Situation Awareness Analysis and Measurement*, pp. 147-174. Mahwah, NJ: Lawrence Erlbaum Associates.

Franke, U. and Brynielsson, J. (2014). Cyber situational awareness – a systematic review of the literature. *Computer Security* (46), pp. 18–31. https://doi.org/10.1016/j. cose.2014.06.008.

Hirvonen, P. (2022a). Expectations And Mindsets Related To GDPR. In T. Eze, N. Khan, & C. Onwubiko (Eds.), *ECCWS 2022: Proceedings of the 21st European Conference on Cyber Warfare and Security* (pp. 360-367). Academic Conferences International Ltd. Proceedings of the European conference on cyber warfare and security, 21. https://doi.org/10.34190/eccws.21.1.238

Hirvonen, P. (2022b). A Review of GDPR Impacts on Information Security. *In PACIS 2022: Proceedings of the 26th Pacific Asia Conference on Information Systems. AI-IS-ASIA*: Artificial Intelligence, Information Systems, in Pacific Asia (Article 83). Association for Information Systems. https://aisel.aisnet.org/pacis2022/83/

Husák, M. et al. (2022). CRUSOE: A toolset for cyber situational awareness and decision support in incident handling. *Computers & Security*, 115, 102609.

Jiang, L., et al. (2022). Systematic Literature Review on Cyber Situational Awareness Visualizations. *IEEE Access.*

KPMG. (2017). O impacto do regulamento geral de protecc¸a˜o de dados em Portugal. https://assets.kpmg/content/dam/kpmg/pt/pdf/pt-2017-rgpd.pdf.

Mahendiran, A. and Appusamy. R. (2018). An intrusion detection system for network security situational awareness using conditional random fields. *Int. J. Intell. Eng. Syst*., (11:3), pp. 196–204.

Panagiotis, T., Nomikos, N., Michailidi et al. (2019). Hybrid Clouds for Data-Intensive, 5G-Enabled IoT Applications: An Overview, Key Issues and Relevant Architecture. *Sensors* (19:16), p. 3591. https://doi.org/10.3390/s19163591.

Presthus, W., Sørum, H. and Andersen, L.R. (2018). GDPR compliance in Norwegian companies. *Norwegian Conference for IT Use in Organisations* (NOKOBIT), pp. 1-15.

Puhakainen, P. (2006). *A design theory for information security awareness.* University of Oulu.

Pöyhönen, J., Nuojua, V., Lehto, M. and Rajamäki, J. (2019). Cyber Situational Awareness and Information Sharing in Critical Infrastructure Organizations. *Information & Security: An International Journal* (43:2), pp. 236-256.

Rajamäki, J. and Katos, V. (2020). Information Sharing Models for Early Warning Systems of Cybersecurity Intelligence. *Information & Security: An International Journal* (46:2), pp. 198-214.

Renaud, K., & Ophoff, J. (2021). A cyber situational awareness model to predict the implementation of cyber security controls and precautions by SMEs. *Organizational Cybersecurity Journal: Practice, Process and People*, 1(1), pp. 24-46.

Rendall, K., Mylonas, A. and Vidalis, S. (2021). Toward situational awareness in threat detection. A survey Edited by: Kim-Kwang Raymond Choo. *Forensic science*. Wires.

Rowan, W., O'Connor, Y., Lynch, L., and Heavin, C. (2021). Comprehension, Perception, and Projection: The Role of Situation Awareness in User Decision Autonomy When Providing eConsent. *Journal of Organizational and End User Computing* (JOEUC), 33(6), pp. 1-31.

Rihoux, B. (2006). Qualitative comparative analysis (QCA) and related systematic comparative methods: Recent advances and remaining challenges for social science research. *International sociology*, 21(5), pp. 679-706.

Sanneman, L. and Shah, J.A. (2020). A Situation Awareness-Based Framework for Design and Evaluation of Explainable AI. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds) *Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. EXTRAAMAS 2020. Lecture Notes in Computer Science (12175). Springer, Cham. https://doi.org/10.1007/978-3-030-51924-7_6.

Saurez, E., Hong, K., Lillethum, D., Ramachandran, U., Ottenwälder, B. (2016). Incremental deployment and migration of geo-distributed situation awareness applications in the fog. *Proceedings of the 10th ACM International Conference*, Irvine, CA, USA, pp. 258–269.

Seo, J., Kim, K., Park, M., Park, M. and Lee, K. (2017). An analysis of economic impact on IoT under GDPR. *8th International Conference on Information and Communication Technology Convergence* (ICTC), Korea, pp. 879-881. https://doi.org/10.1109/ICTC.2017.8190804

Schaberreiter, T. et al. (2022). A cybersecurity situational awareness and information-sharing solution for local public administrations based on advanced big data analysis: the CS-AWARE project. *In Challenges in Cybersecurity and Privacy-the European Research Landscape*, pp. 149-180. River Publishers.

Sirur S., Nurse J. and Webb H. (2018). *Are We There Yet?: Understanding the Challenges Faced in Complying with the General Data Protection Regulation (GDPR).* DOI 10.1145/3267357.3267368.

Tankard, C. (2016). What the GDPR means for businesses. *Network Security*, (6), pp. 5-8.

Teixeira, G., Mira da Silva, M. and Pereira, R. (2019). The critical success factors of GDPR implementation: a systematic literature review. *Digital Policy, Regulation and Governance*, (21:4), pp. 402-418. https://doi-org.ezproxy.jyu.fi/10.1108/DPRG-01-2019-0007

Tikanmäki, I., and Ruoslahti, H. (2019). How Are Situation Picture, Situation Awareness, and Situation Understanding Discussed in Recent Scholarly Literature? In J. Bernardino, A. Salgado, & J. Filipe (Eds.), IC3K 2019 / KMIS 2019: *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, (3), pp. 419-426. SCITEPRESS Science and Technology Publications. https://doi.org/10.5220/0008494104190426.

Toxirjonovich, O. N., and Tulanovna, X. G. (2022). Situational awareness gaps and opportunities for cyber security. *ACADEMICIA: An International Multidisciplinary Research Journal,* 12(1), pp. 512-518.

Vieweg, S., Castillo, C. and Imran, M. (2014). Integrating social media communications into the rapid assessment of sudden onset disasters. *Social Informatics*, pp. 444–461. Springer: Berlin, Germany.

Organisational GDPR Investments and Impacts

Pauliina Hirvonen
University of Jyväskylä, Jyväskylä, Finland

pauliina.a.hirvonen@student.jyu.fi

Abstract: The aim of this empirical multi-case study is to understand the GDPR investments and impacts of the organisations. Among these, the measuring experiences related to GDPR and information security (Isec), and the future expectations are examined. Several interesting findings were recognised, which also enabled further suggestions. First, an understanding of the organisations' investments and their impact is built by gathering information about the actions that organisations made to fulfil the GDPR requirements. In the second phase, it is deemed necessary to examine how organisations experience the measures and evaluation of GDPR development and progress, in order to understand how respondents, end up evaluating the impact of their investments. In the third phase it is considered necessary to consider the future development of GDPR and the challenges and opportunities it brings to organisations, in order to understand how the experiences so far affect preparations for the future. The final phase of evaluation focuses on finding out what impact the GDPR has had on organisations. On the one hand, it is possible that the total investment in the GDPR may also correlate with the development of the organisational Isec maturity, because GDPR has brought more resources and visibility to the organisation's Isec, and operations have become more systematic. On the other hand, organisations with an already high level of Isec maturity and organisations operating in a regulatory-focused industry may accept the GDPR-based Isec investments more easily. If GDPR is tightly integrated with both the organisation's information security and the business functions under the responsibility of executive management, it may support the organisation's business and information security development. This research serves GDPR authorities, organisational executives, persons in charge of GDPR/information security/cybersecurity, service providers and academia.

Keywords: GDPR investments, Empirical GDPR research, Organisational GDPR impacts, Privacy impacts, Information security development, Privacy development assessment

1. Introduction

General Data Protection Regulation (GDPR) brought opportunities in many areas of everyday life for European citizens, technology organisations and different businesses, but also unwanted challenges and interpretation complexities (Tene et al., 2019). As previous research was not able to look at the phenomenon holistically, the aim of this research was to understand comprehensively what GDPR has demanded and caused. The main research question for the multi-case study was: what are the organisational efforts for GDPR and how has GDPR impacted the organisational information security (Isec)? In this research, *the investments concept* refers to all tangible and intangible input, workload, and efforts the organisation has made to meet GDPR requirements. Investment types are treated in this study as qualitative units, and not as precise quantitative measures. *The impacts of GDPR* refer to all the consequences that the regulation has caused for organisations.

The research issue is approached through a multi-case study consisting of eight cases as the eight separate companies. Research aspects included providing several conclusions related to organisations' experiences in GDPR resourcing and measuring organisational Isec development, expectations for future GDPR development and the various impacts of GDPR are made. Suggestions for practices and future research are proposed as well. This research serves GDPR authorities, organisational executives, persons in charge of GDPR/information security/cybersecurity, ICT/ISEC-service providers and academia. The research is organised as follows: Section 2 discusses previous research. The methodology is presented in Section 3. Section 4 includes the analysis. The discussion is in Section 5 and, finally, the conclusions are presented in Section 6.

2. Previous Research

The purpose of reviewing previous research is not to unpack the GDPR articles or requirements in detail, but to understand existing aspects related to organisational GDPR experiences. Perspectives of GDPR efforts and resourcing were found for example by Dibble (2020), who investigated that, instead of being just a once-off exercise, GDPR is more about developing a pervasive framework to cherish the privacy culture. Dode (2018) noted the challenge of correct interpretation and legal interpretation of GDPR requirements. Politou et al. (2018) described the challenge of GDPR for compliance development as containing a restricted number of technical implementation instructions. Tikkinen-Piri et al. (2018) estimated GDPR to require remarkable financial and

human resources and employee training. Degeling et al. (2019) noticed that ICT companies moved in phases to GDPR rather than creating new approaches in their system architecture for GDPR compliance. Dode (2018) mentioned significant requirements for ICT planning and data storage and lineage due to GDPR legal function and compliance. Li et al. (2019) also stated that, organisations must invest, particularly in manpower and resources, to evaluate, develop and redesign their technology platforms and data architectures, update data protection (DP) policies, change advertising policies and adjust data storage and processes.

Most of the research discussed GDPR impacts. Several research (e.g., Poritskiy et al., 2019; Politou et al., 2019; Dellei, 2009) recognised increased confidence and clarifications as the main benefits of GDPR. Poritskiy et al. (2019) identified that organisations experienced hard times with GDPR requirements, especially with the execution of system and process audits and the application of the right to erasure. This survey (Poritskiey et al., 2019) also indicated that the state of implementation of GDPR and the company type are important factors impacting on how GDPR benefits and challenges are perceived. Also the size of the organisation seemed to matter on GDPR impacts (see for example Phillips, 2019; Politou et al., 2019; Bershidsky, 2018 and Goddard, 2017). Li et al. (2019) noted that GDPR complicates the operations of organisations, limits their scope and increases operating costs, which subsequently hampers the productivity and effectiveness of these emerging technologies. Tsohou et al. (2020) noticed that privacy, security, legal and technology acceptance became visible and then integrated GDPR compliance platform requirements. Jia et al. (2018) found negative GDPR impacts on technology investment appearing, particularly for nascent ventures (0 – 3 years) compared to their US counterparts. Impacts are also considered from many other technological aspects (see Wallace and Castro, 2018; Mackay, 2017; Greengard; 2018; and Nautsch et al., 2019).

It is clear that the GDPR is very likely to have an impact on organisational Isec and cyber security (Csec) aspects as GDPR Article 5 imposes regulations on these matters. For example, Mikkelsen et al. (2019) estimated that the investment falls on security controls, continued investments in Csec, and improving internal processes since organisations have to automate and streamline to meet the GDPR requirements. Layton and Elaluf-Claderwood (2019) reported that organisations are faced with remarkable investment costs, especially on privacy compliance. Bilyk (2018) shared ITC companies' challenges to build applications that use personal data in technical and organisational environments. Krikke et al. (2019) explained privacy-by-design to support ICT companies to develop better security solutions, while Fimin (2018) and Bennett (2018) revealed GDPR to improve data management process.

As review shows, the GDPR resourcing and impacts have been examined from several specific perspectives. However, most of the previous research lacked approaching resourcing and impacts comprehensively. All four aspects, including efforts and impacts, measuring the GDPR/Isec development through GDPR efforts, and future expectations of GDPR were not combined in previous research. Empirical research located in Northern Europe was also limited. This research filled the gap by recognising factors related to phenomena based on experiences of eight companies.

3. Methodology and Research Implementation

This research was a multi-case study consisting of eight cases as the eight separate companies and the research aspect as the GDPR impact phenomena examined within each case. Gustafsson (2017) explained the necessity for a multi-case study, when there is more than one single case and, as Eisenhardt and Graebner (2007) described, there is a need for wider exploration of research questions to build more convincing results. The research aspect included four sub-aspects. The research aspect and the main research questions per sub-aspects are presented in Table 1.

Table 1. **Research Aspect and Sub-aspects of the Multi-case Study**

Research aspect	Research sub-aspect	Research question
Organisational GDPR Investments and Impacts	Organisational Investments and Efforts	What has been invested in GDPR in the organisation?
	Measuring Organisational GDPR and Isec Development	How are the development of GDPR /Isec evaluated and what are the main conclusions of the development?
	Future Expectations and Costs	What are future expectations of GDPR and what does it demand from organisations?
	GDPR Impacts	What are the GDPR impacts for the organisation?

Eight interviews were conducted by the author in eight medium or large sized Finnish companies in the autumn in 2021. Each research sub-aspect contained several open-ended interview questions. The industry types were energy, media, engineering, retailing, construction, and software. 1 - 3 participants were interviewed for each organisation. The interviewees worked in varying roles and were persons in charge of GDPR in companies.

Qualitative approach is suitable for research where the examination of the phenomenon is based on subjective experiences without unambiguous truth (Paley, 2000). An advantage of the qualitative multi-case study approach is also that the researcher may refer to the similarities and differences between different cases in a multi-case study, when compared to one case study. This was done by using explanation building techniques together with cross-case comparisons and theming as an analysing method. The data was coded inductively. The thematic analysis increased the relevant content of all results data and supported to find the connections of the interferences of the different approaches in and between cases.

4. Analysis of the Results

The main research question examined what the organisational efforts for GDPR are and how has GDPR impacted the organisational information security (Isec). As noted in Introduction section, the investment types are treated here as qualitative units, and not as precise quantitative measures. The analysis of the results were built in four phases. First, an understanding of the organisations' investments and their impact was built by gathering information about the actions that organisations made to fulfil the GDPR requirements. Next step included examining how organisations measure GDPR and its connection to Isec development. In the third phase the future development of GDPR was considered. The final phase observed what impact the GDPR has had on organisations.

1) Organisational Investment and Efforts for GDPR

GDPR investments were intangible and tangible, direct but also indirect (would have taken place without GDPR). Some investments were non-recurring, for example, system investments, and some were ongoing activities, such as training. *Investment types were identified in seven different areas: management, human resources, finance, organisational change, operational processes, ICT and cooperation*. The results showed that the main effort required human resources (working time). The GDPR was promoted in many organisations in the same way as other development projects, and experts are involved where appropriate. Permissions and resources related to the most significant investments and development were sought from the executive management (EM) , but training and practical compliance were typically decided at the DP organisation level. Three effective means were mentioned to support acceptance of GDPR investment proposals: 1) invoking the content of the regulation, 2) a risk assessment (what would the impact be if there were no change), and 3) an organisational continuity management/development perspective. Decisions related to the fulfilment of the minimum GDPR requirements, decisions that can be economically justified, and decisions that do not incur costs were considered straightforward.

The most challenging decisions mainly concerned upgrades to existing systems or ICT solutions that otherwise contained differences of interpretation. For the most challenging or employable GDPR investment ranged from individual activities to the overall implementation of the GDPR. Four responses emphasised the specific challenges relating to the initial phase, when the content of the requirements was not entirely clear due to a lack of official guidance information. The mapping of the location of personal data, impact assessment, execution of requests for personal data, system investments and "kneading" staff in GDPR matters were also mentioned as the single most challenging aspects. Also, overall compliance and, for example, changing the organisation's operating models and cooperating with partners outside the EU, was seen as confusing according to two companies. The responses revealed differences in the assessment of the significance of the total investment. The total investment was estimated to be important for the organisation (externally reliable player, business development, DP-focused contract expertise management) according to five responses. *The nature of the operations of the organisations also determined how much the investment was seen to benefit the business*. GDPR was mostly seen as an ongoing process rather than a one-time performance.

2) Measuring Organisational GDPR and Isec Development

The impact of GDPR on the specific areas of Isec development and ways to evaluate it in organisations have been varied. All (8/8) companies however agreed that GDPR has had a positive influence on Isec development. The

majority (6/8) of interviewees agreed that GDPR and impact assessment expertise due to training and testing, agreements with partners at various levels of the organisation have brought more resources and visibility to the organisation's Isec and that operations have become more systematic, meaning an increase in Isec training, instruction, testing and preparedness in many organisations. GDPR has improved Isec, but also indirectly Csec and overall security maturity in the organisations 3/8), and intensified cooperation between these functions in organisations (5/8). For two organisations, GDPR and Isec together have enhanced or created a new business opportunity for responsibility. In many organisations (6/8) , DP and Isec were separate operations, although they were implemented in parallel, but *strengthening the dialogue between GDPR, Isec and business perspectives was mentioned as a success factor of organisational development*. Isec has been essential in the construction and maintenance of GDPR compliance, particularly through technical Isec system solutions. In three organisations, Isec is an interpreter between the implementation of GDPR requirements and ICT solutions. Isec and Csec were in some organisations completely reorganised, helped by GDPR. Based on the analysis of the responses it was found that *organisations with a high level of Isec maturity and organisations operating in a regulatory-focused industry were associated with a more flexible acceptance of GDPR-based Isec investments proposals.* However, the development of Isec has, from a GDPR perspective, mainly meant raising awareness of the risks involved in everyday life and operations and updating or acquiring technical Isec solutions. Isec investments (e.g., control) may also still have negative costs for some organisations. In particular, five arguments proved to be essential to GDPR-related Isec investment proposal acceptance (e.g., purchase of a service) in organisations: business continuity, raising the level of the SA of the Isec, risk assessment, importance of reputation and financial threat.

A significant part of GDPR development assessment in every company came from internal feedback and questions, and training evaluations. The interviewees mentioned the following activities: monitoring of DP reporting in internal reporting, evaluation of training and training volumes, analysis of the number and quality of personal data requests, process control, analysis of the number and quality of data breaches, internal contacts, feedback and questions, reviewing DP and Isec status of new parts and functions of the organisation (e.g., after the expansion of the organisation). Organisations may also rely on external auditors to assess Isec development.

3) Future Expectations

With regard to the further development of the GDPR, it was hoped that all the legal demands that organisations should take into account, for example, marketing data, would be covered under one piece of legislation. In addition, timely, unambiguous, detailed and practical industry-specific guidance (e.g., industry-specific checklists) was requested. In particular, guidance on the automation implementation of e-privacy regulation, the formation of DP organisations, and effective tools and clear guidelines for impact assessment were sought. It was hoped that GDPR oversight would take into account the size and industry of organisations and degree of sensitivity to personal information. In addition to fines, it was seen as necessary for the GDPR authorities to provide tools to solve the problems. The specific national issues and the impact of Brexit were seen as a challenge for the future. Experience to date with the GDPR has shown organisations in particular the need to take into account the flexibility of change in all activities of the organisation, the need to identify the location of personal data in the organisation and to take DP into account in the development of new business. The most significant challenges for the future GDPR implementation were data management, contract negotiations for the processing of personal data across EU borders, identification of regulations binding the organisation, delays in guidance issued by regulatory authorities, new solutions requiring evaluation, partnership agreements and changes in the organisation's operations as a result of Brexit.

The most significant development needs of GDPR authorities in the future were the clarification of steering roles and the development of frontloading, improving the clarity and concreteness of instructions, concentrating supervision on GDPR-sensitive organisations and addressing GDPR-induced inequality in the competitive field. The authorities were expected to make the DP management systems available for organisations. In particular, EU-level authorities were requested to have more capacity for a faster cycle of processing decisions, guidelines and recommendations.

4) GDPR Impacts on Organisations

The results showed that the GDPR impacts were generally assessed by most to be significant, pervasive, and lasting. Impacts were seen to be both internal and external (e.g., customers). Some positive impacts are:

responsibilities and accountability, growing importance of partnership agreements, strengthening GDPR awareness and compliance (requirements and guidelines), developing and increasing operating models and processes, access to systems, development of Isec and Csec and overall security guidelines, improved transparency for customers and organisational development of responsibility. The negative effects of GDPR were seen as ongoing resourcing needs (people, money, management) required to maintain compliance in organisations. GDPR was found to be labour intensive for non-personally sensitive organisations. It was noted that bureaucracy related to official guidance increased.

The most significant effects mentioned were the development of the organisation's DP culture and awareness, the benefits for organisational investments in general, as well as the improvement in customer confidence, the decrease in the response time of one's own service and the increase in system knowledge. Although the GDPR process has been a challenging and complex entity for most, organisations managed to evaluate, prepare and adapt to GDPR requirements and to anticipate the effects on organisations, which helped to reduce unexpected or false effects on a wider scale. The small number of requests for personal information or action were the most unexpected positive effects according to four companies. Three interviewees were also surprised by the lack of control and sanction by the official GDPR governing bodies. According to the results, the GDPR did not bring any new competence requirements, but rather changes in small practices. The key GDPR competency requirements related to the following areas:

- Knowledge Of The GDPR Content
- Interpretation And Applications
- Contract Negotiation
- Staff Engagement
- GDPR, Isec And Csec Integration
- Training And Development
- Raising Of Organisational GDPR Situational Awareness (SA)
- Knowledge And Awareness Of Organisation
- Competency Mapping
- Personal Data Process Management
- Project Management

When asked about the overall success of the GDPR, it was mentioned that the goal of the GDPR has been successful and the GDPR was seen to have raised society's DP awareness internationally and Isec expertise of Finnish society. It was mentioned that the GDPR has brought DP cases to industries it had not previously touched on and brought deeper understanding to DP control and the threat of sanctions, but the definition of data breach should be clarified. For the authorities, there was also a need to reduce the response delay and they were encouraged to consider the sector, size and GDPR sensitivity of the organisation. It was mentioned by the majority of the companies (5/8), that the lack of guidance from the official GDPR has had a wide impact on some organisations and that the implementation of the GDPR in organisations would have been more controlled, coherent and streamlined if clear official guidance on requirements had been provided on time.

5. Discussion, Implication for Practices and Restrictions

The result highlighted that the GDPR has required (and still requires) a significant amount of tangible and intangible investments and resources. Totally investments related to seven different areas were identified. GDPR has had a real impact on the development of information security in most of the examined organisations. Some organisations, especially ones with lower maturity levels of Isec, may still find it challenging to justify investment needs to the decision-makers. Organisations with an already high level of Isec maturity and organisations operating in a regulatory-focused industry may accept the GDPR-based Isec investments proposals more easily. However, the practices found to be successful for achieving investment decision approval were shared in this research. As a cross-cutting ongoing process for an organisation, GDPR, if closely integrated with both the organisation's information security and the business functions under the responsibility of EM, may support the organisation's business development and information security development.

Based on the analysis, it is possible that the total investment in the GDPR may also correlate with the development of the organisational Isec maturity and indirectly support the development of Csec and overall security, because GDPR has brought more resources and visibility to the organisation's Isec, and that operations

have become more systematic. As noted in the analysis, GDPR and Isec together have created a new business opportunity for responsibility and it is suggested that organisations enhance the development of GDPR, Isec, Csec and overall security at the same, integrating all these aspects into business functions in every level of the organisations. Practically, this may happen by increasing DP awareness and increasing impact assessment expertise through guidance, training and testing. Also, through agreements with partners at various levels of the organisation, procurement, system upgrades, developments, Isec tools and practices may support the integration. This kind of approach may also support faster integration of information security into all activities of the organisation. However, it seems that organisations still have a long way to go in turning GDPR into business opportunities by themselves.

Related to challenges experienced with the GDPR: legacy systems, system upgrades/integrations or acquisition, ICT and Isec companies could support organisations to balance optimal solutions and also integrate the DP and Isec activities more effectively. To support organisations to deal with future GDPR developments, including development assessment, it is suggested that the GDPR authorities provide certification of standardisation that may help organisations to select appropriate and verified tools, consultants, partners and service providers. The future GDPR challenges relate to the guidance for implementation of automation, e-privacy regulation, the DP organisation development, guidelines and tools for impact assessment and data management, contract negotiations for the data processing across EU borders, partnership agreements, Brexit impact assessment and getting reliable guidance and decisions from officials without delay. GDPR authorities' resources may be used more efficiently by shifting the focus towards the following areas: the clarification of steering roles, the development of frontloading, improving the clarity and concreteness of instructions, concentrating supervision on GDPR-sensitive organisations and addressing GDPR-induced inequality in the competitive field. When comparing these observations of GDPR challenges to research of Poritskiy et al. (2019), two aspects did not come up in the interviews, namely limits the growth of emergent technologies and increasing technical complexity. However, two aspects classified as challenges by Poritskiy et al. (2019), namely employee training and increasing budget associated with security operations, emerged here in the interview as positive or neutral effects.

The major implications in Isec development are the development of the organisation's DP culture and awareness, benefits for the organisational investments in general and improved customer confidence, decrease of response time of own services and increased system knowledge. The data management, opposite to Poritskiy et al. (2018), was seen here as a challenge. As known publicly, there is a huge annual increase in the amount of data as technology advances and expands. For example, only in Finland, between 2019 and 2020, the amount of data increased by about 19% per year (Finnish Federation for Communications and Teleinformatics, 2022). Organisations outside the ICT may consider data management challenging.

These findings can provide interesting insights, especially for official GDPR authorities and law developers, as the results provide observations and recommendations not only from an organisational perspective, but from a more holistic point of view. At the same time, the results serve as authentic feedback on the experiences of medium-sized and large organisations in implementing GDPR during 2018-2021, which can help authorities to improve communication towards organisations and to improve the quality of guidance and information. It is hoped that the results will have a particular impact on organisations' operations and facilitate the implementation of the GDPR and related laws. For EM), the results also show that the organisation's business and its development can successfully be integrated with GDPR, Isec/Csec and the development of overall security, and that they must have an ongoing active and open dialogue in the day-to-day life of the organisation. It may also be useful both for EM and DP organisations to understand the challenges and opportunities associated with GDPR investment. As the results showed the EM's support and commitment to GDPR are critical to achieve GDPR compliance in the organisation. Results may activate and innovate service and solutions providers and ICT, Isec and Csec companies to apply the findings in DP, Isec and Csec product and service design, and support requirement specification and intensify communication with organisations. For external partners, the study will provide detailed information of the content and the timing of the support needs, so the services can be targeted optimally. Research also included suggestions for further academic empirical research, as it touches the surface of several significant GDPR subject areas.

5.1 Restrictions and Future Study

As usual, the most common restriction of the case study research relates to the lack of generalisability of the research results. The research data, consisting of interviews, are subjective, as well as the analysis of the findings, because it is based on the experiences of individuals.

Future research is needed to gain a more comprehensive perspective of this wide research issue. It is recommended to explore the issue by using different research strategies, approaches and methods, but also taking a wider set of samples, and expanding the scope to other EU countries or outside EU borders. Also, as the GDPR landscape is changing, conducting follow-up studies would help to assess long-term effects.

6. Conclusions

A multi-case study consisting of eight cases as the eight Finnish companies was conducted to explore the GDPR experiences in organisations. This research complemented existing and found new perspectives for the issue. The research questions of the multi-case study explored what GDPR has demanded and caused for organisations. The answers to all four research sub-aspects were found and several inferences related to organisations' experiences of GDPR resourcing, measuring GDPR and Isec development, expectations of the future development of the GDPR and the variable implications of GDPR were made. Investments related to seven different areas were identified. It is possible that the total investment in the GDPR may also correlate with the development of the organisational Isec maturity, because GDPR has brought more resources and visibility to the organisation's Isec, and operations have become more systematic. However, organisations with an already high Isec level of maturity, and organisations operating in a regulatory-focused industry, may accept the GDPR-based Isec investments more easily. As a cross-cutting ongoing process for an organisation, GDPR, if closely integrated with both the organisation's information security and the business functions under the responsibility of EM, supports the organisation's business development and information security development. Besides several positive effects, the initial phase of implementation in particular was perceived as very challenging due to unclear communication from the GDPR authorities. Research included suggestions for practices and future research. This research is hoped to serve several parties working with GDPR: authorities, organisational executives, persons in charge of GDPR, ICT/ISEC service providers and academia.

References

Bennett, C. (2018). The European general data protection regulation: an instrument for the globalization of privacy standards? *Information Policy*, 23(2), pp. 239-246.

Bershidsky, L. (2018). *Europe's privacy rules are having unintended consequences.* www.bloomberg.com/opinion/articles/2018-11-14/facebook-and-google-aren-t-hurt-by-gdpr-but-smaller-firms-are

Bilyk, V. (2018). *Challenges and benefits of GDPR implementation.* https://theappsolutions. com/blog/development/gdpr-challenges-and-benefits/

Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., & Holz, T. (2019). We value your privacy...now take some cookies: measuring the GDPR's impact on web privacy. *Proceedings of the Network and Distributed Systems Security (NDSS) Symposium 2019*, San Diego, CA, pp. 1-15.

Dellei, L. (2019). *GDPR compliance as a competitive advantage.* www.isaca.org/Knowledge-Center/Blog/Lists/Posts/Post.aspx?List = ef7cbc6d-9997-4b62-96a4-a36fb7e171af&ID=1133&utm_referrer=direct%2Fnot%20provided

Dibble, S. (2020). The GDPR Marathon. *Solicitors journal,* 163(3), p. 40.

Dode, A. (2018). The challenges of implementing General Data Protection Law (GDPR). *14th International Conference in "Standardization, protypes and quality: a means of Balkan countries' collaboration*", September 21 - 22, 2018, Tirana, Albania. https://www.researchgate.net/profile/Albi-Dode/publication/327829348_The_challenges_of_implementing_General_Data_Protection_Law_GDPR/links/5ba74542299bf13e6045fca9/The-challenges-of-implementing-General-Data-Protection-Law-GDPR.pdf

Eisenhardt, K. M., & Graebner, M. E. (2007). Theory building from cases: Opportunities and challenges. *The Academy of Management Journal,* (50:1), pp. 25-32.

Finnish Federation for Communications and Teleinformatics. (2022). Matkaviestinverkossa siirretty data.https://www.ficom.fi/ict-ala/tietopankki/viestintaverkot-tietopankki/kiintea-ja-mobiililaajakaista/matkaviestinverkossa-siirretty-dat/

Fimin, M. (2018). *Five benefits GDPR compliance will bring to your business*. www.forbes.com/sites/forbestechcouncil/2018/03/29/five-benefits-gdpr-compliance-will-bring-to-your business/#1d2981cb482f

Goddard, M. (2017). The EU General Data Protection Regulation (GDPR): European Regulation that has a Global Impact. *International Journal of Market Research*, (59:6), pp. 703–705. https://doi.org/10.2501/IJMR-2017-050.

Greengard, S. (2018). Weighing the Impact of GDPR. *Communications of the ACM* (61:11). Society. News. DOI:10.1145/3276744

Gustafsson, J. (2017). *Single case studies vs. multiple case studies: A comparative study.*

Li, H., Yu, L., and He, W. (2019). The Impact of GDPR on Global Technology Development. *Journal of Global Information Technology Management*, (22:1), pp. 1-6. DOI: 10.1080/1097198X.2019.1569186

Jia, J., Jin, G. Z., and Wagman, L. (2018). *The short-run effects of GDPR on technology venture investment* (No. w25248). National Bureau of Economic Research.

Krikke, J., Valgaeren, E., and Origer, G. (2019). *Complying with the general data protection regulation* (GDPR). NBER Working Paper No. 25248. November 2018. JEL No. D43,D8,L13,L15,L5. www.stibbe.com/en/expertise/practiceareas/data-protection/general-dataprotection regulation/what-are-the-challenges

Layton, R., and Elaluf-Calderwood, S. (2019). A Social Economic Analysis of the Impact of GDPR on Security and Privacy Practices. *12th CMI Conference on Cybersecurity and Privacy* (CMI), pp. 1-6. doi: 10.1109/CMI48017.2019.8962288.

Mackay, D. (2017). *The impact of GDPR from a technology perspective – is your platform ready?* https://www.ness.com/11101-2/

Mikkelsen, D., Soller, H., Strandell-Jansson, M., and Wahlers, M. (2019). *GDPR compliance since May 2018: A continuing challenge*. Risk Practice. McKinsey & Company. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/Risk/Our%20Insights/GDPR%20compliance%20after%20May%202018%20A%20continuing%20challenge/GDPR-compliance-since-May-2018-A-continuing-challenge.pdf

Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I., & Evans, N. (2019). The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps towards a Common Understanding. Proc. *Interspeech 2019*. https://doi.org/10.48550/arXiv.1907.03458

Paley, J. M. A. (2000). Paradigms and presuppositions: The difference between qualitative and quantitative research. *Scholarly Inquiry for Nursing Practice,* (14:2), pp. 143-155. https://www.proquest.com/scholarly-journals/paradigms-presuppositions-difference-between/docview/893000198/se-2?accountid=11774

Phillips, N. (2019). Keep it: Maintaining competition in the privacy debate. Remarks for Internet Governance Forum.

Politou, E., Alepis, E., and Patsakis, C. (2018). Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions. *Journal of Cybersecurity*, (4:1). https://doi.org/10.1093/cybsec/tyy001

Politou, E., Michota, A., Alepis, E., Pocs, M., and Patsakis, C. (2019). Backups and the right to be forgotten in the GDPR: an uneasy relationship. *Computer Law & Security Review,* (34:6), pp. 1247-1257.

Poritskiy, N., Oliveira, F., and Almeida, F. (2019). The benefits and challenges of general data protection regulation for the information technology sector. *Digital Policy, Regulation and Governance*, (21:5), pp. 510-524. doi:http://dx.doi.org/10.1108/DPRG-05-2019-0039

Tene, O., Evans, K., Gencarelli, B., Maldoff, G., and Zanfir-Fortuna, G. (2019). GDPR at Year One: Enter the Designers and Engineers. *IEEE Security & Privacy*, (17), pp. 7-9. 10.1109/MSEC.2019.2938196.

Tikkinen-Piri, C., Rohunen, A., and Markkula, J. (2018). EU General Data Protection Regulation: Changes and implications for personal data collecting companies*. Computer Law & Security Review*, (34:1), pp. 134-153.

Tsohou, A., Magkos, E., Mouratidis, H., et al. (2020). Privacy, security, legal and technology acceptance elicited and consolidated requirements for a GDPR compliance platform. *Information & Computer Security.*

Wallace, N., & Castro, D. (2018). *The impact of the EU's new data protection regulation on AI.* www.datainnovation.org/2018/03/the-impact-of-the-eus-new-data-protection-regulationon-ai/

A Whole-of-Society Approach to Organise for Offensive Cyberspace Operations: The Case of the Smart State Sweden

Gazmend Huskaj[1,2] and Stefan Axelsson[1]
[1]Department of Computer and Systems Sciences, Stockholm University, Kista, Sweden
[2]Geneva Centre for Security Policy (GCSP), Geneva, Switzerland
g.huskaj@gcsp.ch

Abstract: Threat actors conduct offensive cyberspace operations for many purposes, such as espionage, to destroy information assets, and cybercrime. These operations are possible thanks to the innovation and development of information and communications technologies (ICT). Interconnected information systems have transformed societies positively. However, specific states exploit these systems' vulnerabilities to advance their strategic national interests. Therefore, it is important to know how a state can organise itself to defend against threat actors. The purpose of this research is to present how the smart state Sweden can through a whole-of-society approach organise for Offensive Cyberspace Operations. The intent is to conduct an active and independent foreign-, security- and defence policy, but also as a base for deterrence and defence. This article is based on a mixed methods approach. It uses the case study research strategy to discover new information. Fourteen men and women participated in individual semi-structured interviews. The respondents ranged in age from 40 to 65 with more than 20 years of experience in cyberspace operations, intelligence operations, military operations, special forces operations, and knowledge and understanding about information warfare and information operations. The analytic strategies include thematic analysis and quantitative methods to interpret the data. The results show many themes, but the article is especially focused on the themes of Operations, Capability, Policy & Governance, and Legal Frameworks. Finally, a conceptual map of a whole-of-society approach to organise for offensive cyberspace operations is presented inferred from the themes, codes, and content, and mapped to each responsible agency based on the interviews and codes. The answer to the research question is that Sweden should have a whole-of-society approach to organise for Offensive Cyberspace Operations to project power in and through cyberspace with the intent to conduct an active and independent foreign, security and defence policy and for deterrence, as described in Figure 2.

Keywords: Deterrence, Information Systems, Cyberspace capabilities, Offensive cyberspace operations, Smart State.

1. Introduction

Threat actors conduct offensive cyberspace operations for many purposes, such as espionage, to destroy information assets, and cybercrime. These operations are possible thanks to the innovation and development of information and communications technologies (ICT). Interconnected information systems have transformed societies positively. However, specific states exploit these systems' vulnerabilities to advance their strategic national interests. Examples include suppressing their population to ensure their autocratic rule [Ernst and Lee, 2021], stealing knowledge in private companies and industry [PwC and BAE Systems, 2017], and for intelligence purposes in preparation to take over another country's territory [Deibert et al., 2012, Dickinson, 2021]. Once the tanks are invading, the organisations of these states or their proxies continue to conduct offensive cyberspace operations for intelligence purposes and support the kinetic effort [Microsoft, 2022].

Cyberspace, like air, land, sea and space, has been declared a domain for military operations Command. The United States Joint Chiefs of Staff in 2004 declared it as a "'domain' of conflict alongside the air, land, sea, and space domains, and noted DoD must maintain its ability to defend against and to engage enemy actors in this new domain" (US CYBERCOM). Defending and engaging threat actors in and through cyberspace is not new. This idea was based on early discussions as early as the 1970s about computer security. In 1975, Rein Turn and W. H. Ware presented a research paper titled "Privacy and Security in Computer Systems." They argued that "the vulnerability of computerised information has prompted measures to protect both the rights of individual subjects and the confidentiality of research database" [Turn and Ware, 1975]. However, public discussions did not occur until 1983, when the movie WarGames showed "computer hacking" and its implications on national security. The result was the first U.S. Presidential Directive on computer security. Since then, computer hacking, or offensive cyberspace operations, has increased (c.f. the Cyber Operations Tracker or the CyberPeace Institute).

Researchers have identified and studied how to organise for offensive cyber operations, but this research is limited. Devanny et al. [2022] note that cyberspace operations are conducted and discuss these within the concept of "cyber power" and how states should organise themselves, but view it from the angle of a "coherent cyber strategy." Aschmann et al. [2017] discuss how threat actors can plan, prepare and "organise specific strikes." Rohret [2010] discuss how a framework may be used to organise offensive operations. Finally, Kaplan [2017] notes that the U.S. pondered where to organise the new combatant command for offensive cyberspace

operations. During those discussions, the U.S. identified the National Security Agency (NSA) as the best-suited organisation. According to Kaplan, the motive was that the NSA were best equipped to conduct offensive cyberspace operations and because the U.S. was not big enough to have multiple institutions for cyberspace operations. Therefore, to use resources wisely, the conclusion was to establish the United States Cyber Command on "top" of the NSA [Kaplan, 2017].

Identifying how a state should organise itself for offensive cyberspace operations is challenging. Kaplan [2017] is the only author who gives some insights on how policymakers discuss how to organise for offensive cyberspace operations and other researchers (e.g., van Haaster, Willett, Hunker) discuss cyber power. Our research question is, therefore: how can Sweden, a smart state, through a whole-of-society approach, organise for Offensive Cyberspace Operations? The intent is to conduct an active and independent foreign-, security- and defence policy, but also as a base for deterrence and defence. The notion of a "Smart State" is a state that "seek[s] influence based on the power of ideas" (Sundelius, 1995, as cited in Eriksson, 2019).

The starting point for this research article is identifying a framework to measure national power linked to the rapid technological and social changes Tellis et al. [2000a]. The framework provides the base for semi-structured interviews with experienced security professionals in Sweden's intelligence, security and defence sectors. The interviews were analysed through thematic content analysis. The theme size is based on a quantitative analysis of the codes. Based on the interviews and the themes, a conceptual map of a whole-of-society approach to organise for offensive cyberspace operations is described. These methodological considerations are reported in section 3, and the analysis in section 4. Section 5 presents the conceptual map of a whole-of-society approach to organise for offensive cyberspace operations, and the article ends with conclusions and future research. The contribution of this article is a conceptual map of a whole-of-society approach to how Sweden, with its agencies, private sector and academia, can organise itself for offensive cyberspace operations.

2. The Framework to Measure National Power

Rapid technological and social changes affecting national power and the balance of power are nothing new. Huskaj [2023] presents how nations considered defending the sea trade routes, leading to sea power; land power was about defending trade routes on railways; airpower became a concept with the discovery of aircraft. Ampuja and Koivisto [2014] labelled concepts and note shifts from one "time" to another "time", such as from the "post-industrial"-times to the "network society"-times. They argue that information technologies accelerate and enhance the process of information. Major companies and states that understood that having effective infrastructure could increase their competitiveness internationally [Ampuja and Koivisto, 2014].

These shifts, rapid technological and social changes, and how they affect national power, the balance of power, and the nature of warfare, concerned the U.S. Office of Deputy Chief of Staff for Intelligence. More specifically, the concerns that triggered this were:

1. A growing unease with the current aggregate measures of national power used within the intelligence community.
2. A growing suspicion that the nature of warfare itself may be changing in fundamental ways.
3. An increasing concern that the lack of an adequate methodology to assess national power might cause the United States to miss or misinterpret incipient changes in power capability that may be taking place within many countries in the international system.

[Tellis et al., 2000b, p. 1].

The result was the framework to measure national power. It consists of three levels: the national resources level, the national performance level and the military capability level. However, this research study aimed to identify how Sweden should organise itself for offensive cyberspace operations, not to measure Swedish national cyber power. Therefore, the framework was adapted to focus on organising for offensive cyberspace operations.

3. Research Design

The researchers base this research on interpretivism. Humans make meanings based on their social context, time, place, expertise, training and education. Therefore, to create a new and rich understanding of how Sweden can organise itself for offensive cyberspace operations, it was evident to approach very experienced Humans on the topic of this research article. GH approached and asked very experienced people if they wanted to be part of this research. Those were either part of GH's network or recommended by the network.

Fourteen men and women participated in individual semi-structured interviews. The respondents ranged in age from 40 to 65 with more than 25 years of experience in cyberspace operations, intelligence operations, military operations, special forces operations, and knowledge and understanding of information warfare and information operations. Semi-structured interviews lasting between one and two hours enabled data collection. All interviews were conducted and moderated by GH. The researcher presented the cases in Table 2 to the respondents to drive the discussion because [Mandel, 2017] provided extensive detail. He categorises the cases by case name, type of attack, who initiated the attack, and whether the initiator had state support. Although his book provides more details about each case, the respondents are, and were, well aware of each case. Therefore, visualising the cases in this format stimulated the discussion during the semi-structured interviews.

Thematic analysis, as described by Braun and Clarke [2006], is the applied analytic strategy. Thematic analysis is a qualitative analytic strategy that looks for "themes or patterns" [Braun and Clarke, 2006, p. 77] in the data set. The researcher is actively looking for themes or patterns in the dataset by generating a list of codes. These codes, in turn, are analysed and together form a theme broader than the unit of analysis (code) [Braun and Clarke, 2006]. Table 1 presents data extracted from the data set and its coding, followed by how the codes are inferred into a theme: Operations.

Table 1: Data extract with codes applied and resulting theme.

Data Extract	Coded for	Theme
Then you need offensive capability so that you can, in those cases, which many times are about shutting down cyber infrastructure, shutting it down, best, because it costs money to destroy a bunch of servers, costs money and time.	offensiveOperations	Operations
Of those I know in more detail, I would say that I would not use the term enhance, but support. Support to the operations, or to the campaign.	supportOperations	

These qualitative techniques and procedures to interpret the data result in 'a relatively systematic and comprehensive summary or overview of the dataset as a whole' [Wilkinson, 1997, 170]. Thus, applying these techniques and procedures results in a long list of codes. These codes are, as mentioned above, inferred into themes. The result of coding and inferring is a thematic map of 25 themes.

Table 2: The cases presented to the respondents. (Source: [Mandel, 2017]).

Case	DataCorruptionErasure	DataCorruptionviaWormInsertion	DataTheeft	DenialofService	DisruptingPhones	InsertVirus	MalwareInsertion	Defacement	ViolencePresence	InitiatorType	StateSupport
Estonia				X					X	NSG/Gov	Russia
Gaza				X				X	X	Terror Group	N/A
Georgia			X	X				X	X	NSG	Russia
Saudi Aramco	X					X				NSG	Iran
Sony							X			NSG	North Korea
Stuxnet		X								Govs	N/A
Ukraine				X	X			X	X	NSG	Russia

4. Analysis - A Whole-of-Society Approach to Organise for Offensive Cyberspace Operations

This section summarises the major themes of the thematic content analysis, illustrating them with quotes from the respondents. Although figure 1 presents all identified themes, this section will cover only the themes of Operations, Capability, Policy & Governance, and Legal Frameworks.

4.1 Offensive Cyberspace Operations

With its 61 codes, the operations theme is the most prominent theme. The theme has multiple sub-themes on cyberspace operations. The first sub-theme is about operations in general. Here, the respondents talk about offensive and defensive cyberspace operations and that offensive cyberspace operations are not the same as information operations. For example, one respondent notes that an offensive cyberspace capability should be able to conduct "all types [of operations], affect CIA [Confidentiality, Integrity and Availability], steal information, manipulate, and to turn off availability." Another respondent notes that an offensive cyberspace capability should be able to conduct "Attacks on CNI, [and] one must develop the capability to reach as deep as possible on CNI or make it possible that one can do these kinds of attacks, to turn off the power on whoever is threatening us." CNI refers to "Critical National Infrastructure", such as the financial, transportation, and energy sectors.

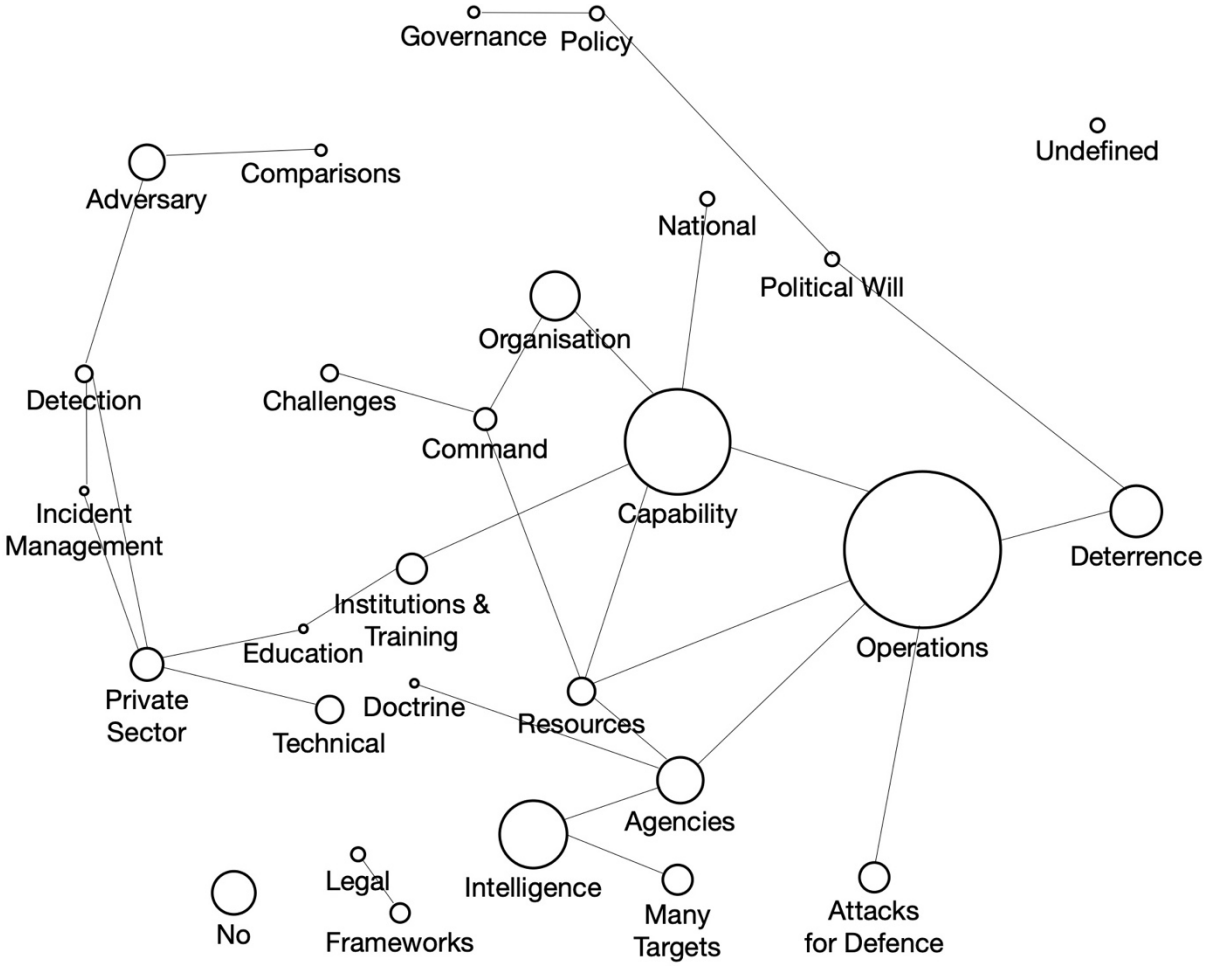


Figure 1: The Developed Thematic Map. Theme size is based on the number of codes. For example, the operations-theme has 61 codes.

In conducting defensive cyberspace operations, a unit must be able to conduct offensive operations. In other words, to defend, one must attack. Four respondents note that to defend; one must be able to attack: to "have an offensive capability to attack for defensive purposes."

One respondent notes that offensive cyberspace operations are not information operations. The respondent notes that "it is interesting to talk about domains. If Russia is trying to affect the election in the US or Sweden,

one should not consider it equal to they [Russia] have conducted a cyber operation. The domains go into each other, but it is a good example to keep them separated. Think of a Venn diagram with a joint space."

The second sub-theme concerns how offensive cyberspace operations should be integrated into information and multi-domain operations. Here, one respondent notes having the ability to conduct offensive cyberspace operations as part of an overall information operation. An example could be to "Combine an information operation with a cyber attack targeting a water cleaning facility to ensure more bacteria in the water could lead to a very effective part." Integration in multi-domain operations is about integrating cyber into the other domains: air, land, sea and space. The purpose is to exploit cyber to the maximum to achieve advantages in other domains.

The third sub-theme is about how to use offensive cyberspace operations. Some examples include influencing the adversary for command-and-control warfare, defensive purposes, sabotage, as a force multiplier, as a support function, for deception, competitive advantages, and to generate cognitive effects. One respondent notes that it is possible to influence the adversary by conducting cyberspace operations and "see" parts or the whole of the adversary's situation awareness: "If you can influence and interfere with the adversary's situation awareness, you can gain an advantage." The respondent counties:

"Whatever adversary information system you can access and interfere or destroy leads to throw the adversary off balance, such as to turn off telecommunication services, payment functions or other parts of communication of the societal structure."

Command and control warfare uses offensive cyberspace operations to impact command and control. One respondent notes:

"You depend on information to conduct command and control. If you limit information through cyber operations or through secondary effects, by not striking directly on the weapon systems, but on supply chains. You are not destroying the weapon system but ensuring [the adversary] does not receive the weapons or fuel."

Offensive cyberspace operations are also a support function. Some examples include supporting other operations, companies in critical national infrastructure, and police investigations.

The fourth sub-theme is about risks related to offensive cyberspace operations. Some examples include unintended secondary and tertiary effects (=cascading effects) and risks leading to a diplomatic crisis. Conducting offensive cyberspace operations that do not lead to secondary and tertiary effects is essential. Should that occur, there are risks the adversary may realise they are under attack and could retaliate: this cannot happen.

4.2 The Cabinet Office and the Policy Level

The respondents noted that an effective cyber capability requires some form of governance, and the governance model differs if operations are for offensive, defensive or collection purposes. The highest national level, the Cabinet Office and the policy level govern the use of offensive cyberspace operations. Governing the use of offensive cyberspace operations requires, according to one respondent, "someone" on the policy level that can "make decisions, [understands] legal frameworks, [with] a legal position on international law, linked to legal frameworks." Another respondent noted that the same person must have "a basic education in college; through that understanding and experience, one can understand the operational parts and then work on the policy level." A third respondent emphasises, "I do not believe you can only educate policy people to understand these things. Things can turn bad. Being at the policy level means being on the strategic level, which means a person on that level is senior: There are no shortcuts in the cyber-area."

The implication is that the "someone" working on the policy level must at least have a primary education on the college level and then get experience with cyberspace capabilities. The road to the policy level is long, and being on that policy/strategic level also means an individual is of senior rank. The "there are no shortcuts in the cyber-area" implies that an individual must have this "basic college education" and, through working experience, understand cyberspace capabilities' positive and negative effects: the highest level must always answer to the political level if something goes wrong. Therefore, anyone working with offensive cyberspace operations on the policy level needs an understanding of the entire chain: from the target, the offensive capability chosen for that target, potential secondary and tertiary effects (also known as cascading effects), legal aspects, to potential risks. Although this may sound complicated and new, it is not. Similar experiences exist in Sweden: the Special Operations Forces (SoF) organisation.

Special Operations Forces Command (SOFC) directs the Special Operations Forces. The strategic level must understand how to leverage the Special Operations Forces to achieve strategic and security policy goals and this capability's strengths and weaknesses. The history of the Swedish Special Operations Forces and the challenges it had to overcome to become the organisation it is today were many (and will not be covered here). Those with insights into that journey can provide examples to extrapolate the cyber-domain. One respondent notes that SF operators "cannot get caught because the politicians on the strategic level will react." To mitigate this, the respondents note that a closed community should know about ongoing SoF operations, but a general community should know that the SoF-capability exists.

4.3 Capability

Capability is about having the power to do Etymonline [2023]. Therefore, cyber capability is about having the necessary institutions, policy, doctrine, training, education and organisation to have the power to conduct cyberspace operations. A capability to conduct cyberspace operations needs to be agile. The necessary institutions range from the policy level down to the individual. The policy level should have experienced people who are senior and understand the technology level and related strengths and weaknesses. The individual level is related to the individual's role concerning his/her expertise within the unit.

The policy, technology, legal and operations domains require their respective institution to provide training and education. This is because the strategic, operational, and tactical levels differ in scope. One respondent notes that "in addition to the unit itself, the whole training system: school, academia, recruitment, training facilities, training and testing will require much coordination. Then the organisation needs its capability for competence development." Another respondent notes:

"You need technical, organisational skills, a unit, with the required capability. Then you need a regulatory framework, legal position, international law linked to the regulatory framework, and rules for peace and war. If you have those two, then you need decision-making mechanisms to evaluate a CONOPS for a cyber operation, take a position, OPSEC, and then be able to make decisions, follow up on implementation, and effects."

The cyber-organisation can have a command-like structure for offensive and defensive cyberspace operations. One respondent notes that "the command-like structure should protect the Armed Forces' systems. The armed forces and the National Defence Radio Establishment (FRA) should develop and organise the capability. It is close to SIGINT; those resources should probably be pooled. We have Network Collection and Offensive capabilities, capability-wise there is not much difference." The unit should consist of "typical roles: infrastructure specialists, intelligence analysts, a lawyer at the tactical level to give direction based on current legislation [that] 'we have the following possibilities.' Otherwise, the whole system fails." Part of the institutional structure is the role of the private sector.

The cyber security private sector has a role, primarily a defensive one. The private sector can conduct red-team exercises and be "posted" in different departments or teams. Giving them additional context, they can become "a reserve organisation like, for example, the British Army LIAG, or like the Estonian Cyber Home Guard, supporting other things than just offensive." LIAG is the British Army Land Information Assurance Group "which comprises professionals selected for their professional skills. This reduces the training burden on the military such that only military skills need to be provided to these volunteers - their technical skills are the main selection criteria" Military Wiki [2023]. LIAG is under the Operational Command of the Joint Cyber Reserve Force [Gov.uk, 2023].

4.4 Legal Frameworks

The importance of regulatory frameworks, legal positions and international law has been mentioned above. Another respondent notes, "There is no legal expertise, as a field, it is less developed than, e.g., cluster bombs and risk distances to the civilian population." The respondent notes, "an operational, legal function that can evaluate [legal frameworks] is important."

5. A conceptual map of a whole-of-society approach to organise for offensive cyberspace operations

The conceptual map results from the researchers' interpretation of the data, coding and inferring, and is based on the research philosophy of interpretivism. It is important to note that different researchers may conclude differently. However, the researchers of this article have extensive professional and academic experience in the

Swedish security system. The combination of interpreting and inferring the collected data and professional and academic experiences justifies this analysis.

Sweden should have a whole-of-society approach to organise for Offensive Cyberspace Operations to project power in and through cyberspace to conduct an active and independent foreign, security and defence policy and for deterrence, as described in Figure 2. The figure results from conceptualisation and inference, as mentioned in Section 3. Each theme is associated with an icon or figure belonging to an agency responsible for that theme (for example, FRA, MPF, SÄPO) or a generic icon or figure such as academia, private sector & industry, and capabilities.

These qualitative techniques and procedures to interpret the data result in 'a relatively systematic and comprehensive summary or overview of the dataset as a whole' [Wilkinson, 1997, 170]. Thus, applying these techniques and procedures results in a long list of codes. These codes are, as mentioned above, inferred into themes. The result of coding and inferring is a thematic map of 25 themes.

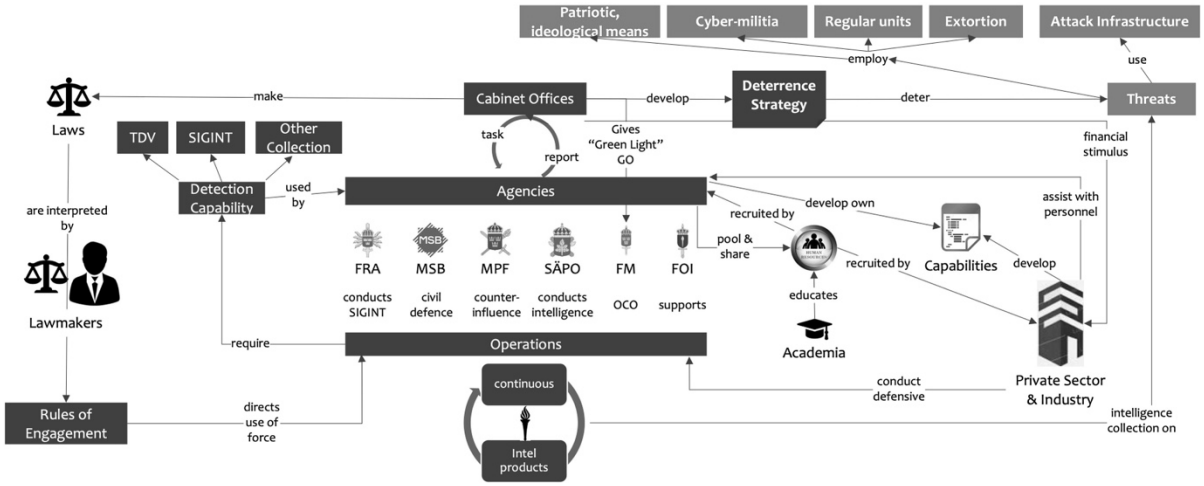


Figure 2: The conceptual and inferred map of all themes, codes, and content, mapped to each responsible agency based on the interviews and codes.

This section describes each concept. Threats use or provide attack infrastructure to their entities, employ patriotic, ideological means, and use extortion to recruit human resources. These entities can take the form of cyber-militia and regular units. A deterrence strategy is required, developed by the Cabinet Offices to deter the threats. Furthermore, an effective cyber capability requires engaging all national resources. This begins at the top level, at the Cabinet Office, by establishing a vision for the cyber capability. This vision also presents that the Cabinet Offices are the ultimate authority for offensive cyberspace operations. Next, the Cabinet Offices share that vision with the agencies, academia, private sector and industry. The vision can take the form of a strategy, such as a cyber security strategy, with discussions on offensive and defensive measures.

The Cabinet Office also makes laws interpreted by lawmakers that break them down into Rules of Engagement (RoE). The RoE's direct use of force for offensive cyberspace operations. The Cabinet Offices hold command over the cyber capability and are responsible for giving FM the "green light" to conduct offensive cyberspace operations.

The Cabinet Offices enable and direct how agencies should work through instructions to the agencies. For the agencies, it is essential to adhere to the Cabinet Offices' instructions and note the differences between these and various principles. The principle of responsibility, while giving good guidance, should receive the attention it requires in the form that it is currently having a fragmenting effect: regardless of how many resources the various agencies receive, the risk is that these are unsatisfactory.

Next, organisation, pooling and sharing resources are critical enablers for the organisation of an effective cyber capability. Organising the capability by co-locating it with other agencies with similar capabilities is an excellent first step to ensure that human and infrastructure resources are wisely managed.

The Cabinet Office provides financial stimulus to the private sector & industry to develop capabilities for the agencies and to pool and share Human Resources. The private sector & industry conduct defensive operations and share threat intelligence with the agencies to increase situational awareness. Additionally, the private sector

and industry can conduct defensive cyberspace operations to protect their own infrastructure. Furthermore, they can support FM, FRA, MPF, MSB and SÄPO with intelligence and software/hardware capabilities.

The private sector and industry can also conduct offensive cyberspace operations when the Cabinet Offices require them to do so from government-related infrastructure. This clearly defines that the smart state is in charge, and one possible scenario when this would happen is when the smart state requires to conduct offensive cyberspace operations. However, the key competencies are within the private sector & industry.

This type of integration between the private and public sectors for defensive and offensive operations is unheard of in the pre-cyberspace era: cyber operators with the required skills to conduct offensive cyberspace operations do not grow on trees, and many different skill sets are needed. For example, training military operators to operate a main battle tank can take up to 7 months, while training a cyber operator for offensive operations can take years.

The National Defence Radio Establishment (FRA) conducts cyber intelligence, surveillance, and reconnaissance (ISR) operations on the threats. They use this intelligence for future cyber intelligence operations and provide intelligence support to other agencies. MSB conducts defensive cyberspace operations. The psychological defence agency (MPF) is responsible for information operations and counter influence-operations.

SÄPO conducts intelligence cyberspace operations on internal threats.

FM conducts offensive cyberspace operations, with support from other agencies, but primarily from FRA. FM targets a threat's attack infrastructure and other cyberinfrastructure. Other targets include telecommunications services, communications channels, communications to air, sea or other societal functions. Additional targets include supply chains, command and control systems, critical national infrastructure, and payment functions. Conducting pure defensive cyberspace operations will probably lead to failed deterrence.

FOI provide support for training and education through their CRATE environment and research. At the same time, academia educates and trains human resources that, in turn, are recruited by the agencies, the private sector & industry. The agencies pool and share the recruited human resources to mitigate the lack of human resources: there must be more competencies in cyberspace operations. Pooling, sharing, and maximising the use of resources requires co-locating the cyber capability at FRA. The nature of offensive cyberspace operations and SIGINT operations are very similar. Co-location also enables a quick deconfliction mechanism and short decision-making lines.

Co-location is essential because offensive cyberspace operations and intelligence collection operations differ from a risk perspective: the military commander takes some risk executing offensive operations when generating effects; the commander conducting intelligence collection may not accept any risk because those operations should go unnoticed. This setup makes it possible to deconflict time-critical operations quickly.

Another way of organising could be as in the USCYBERCOM case: the Director of the NSA is also the Director of US CYBERCOM, a dual-hatted position. The USCYBERCOM case has, until the writing of this article, showed very positive results with empirical evidence such as Operation Glowing Symphony [Darknet Diaries, 2019] and 28 so-called Hunt-Forward Operations, a part of "CYBERCOM's 'persistent engagement,' an effort aimed at proactively defending the U.S. against malicious cyber activity [U.S. Cyber Command, 2022].

6. Conclusions & Future Research

The study builds on an adapted framework for describing national power and extends it by applying it in the cyber domain. Based on the cases, the offensive cyberspace operations, and their impact, the respondents agree there is a need to organise for offensive cyberspace operations. However, some challenges exist, such as needing more human resources with the necessary skills. However, public-private partnerships can overcome these challenges.

The answer to the research question is that Sweden, the smart state, should have a whole-of-society approach to organise for Offensive Cyberspace Operations to project power in and through cyberspace with the intent to conduct an active and independent foreign, security and defence policy and for deterrence, as described in Figure 2. Co-locating the cyber capability at FRA enables the pooling and sharing of Human Resources and maximising the use of other resources. Furthermore, co-location enables the ability to deconflict time-critical operations between the offensive and intelligence roles, which are critical for mission success.

With support from the private sector, the national cyber capability conducts defensive and offensive operations at an age where private companies are leading the technological development, which in a pre-cyberspace era were led by national military establishments and related supporting industries.

Multiple areas for further research exist. One area is to give further insights into how the private sector could be engaged in offensive operations at the government's will. What problems may arise, and how can these problems be solved?

7. Glossary

AI:	Artificial Intelligence.	MISP:	Malware Information Sharing Platform.
AV:	Anti-Virus.	ML:	Machine Learning.
C2:	Command & Control.	MSB:	Swedish Civil Contingencies Agency (Myndigheten för Samhällsksydd och beredskap)
CNI:	Critical National Infrastructure.	OPS:	Operations
CSA:	Cyber Situational Awareness.	OSI Stack:	The Open Systems Interconnection Stack/Model.
DDoS:	Distributed Denial of Service.	OSINT:	Open Sources Intelligence.
FM:	The Swedish Armed Forces (Försvarsmakten)	PFP:	Partnership for Peace.
FRA:	The National Defence Radio Establishment (Försvarets radioanstalt)	PSYOPS:	Psychological Operations.
GCHQ:	Government Communications Headquarters.	SÄPO:	Swedish Security Service (Säkerhetspolisen).
IDS:	Internal Detection System.	SIGINT:	Signals Intelligence.
IOC:	Indicators of Compromise.	TDV:	Technical Detection- and Warning system (Tekniskt detekterings- och varningssystem).
IPB:	Intelligence Preparation of the Battlefield.	USCYBERCOM:	United States Cyber Command.
IPOE:	Intelligence Preparation of the Operational Environment.		

References

CyberPeace Institute. (2023). Cyber threats. Retrieved from https://cyberconflicts.cyberpeaceinstitute.org/threats.

R. J. Agard. AFSC 1B4X1 CYBER WARFARE OPERATIONS CAREER FIELD EDUCATION AND TRAINING PLAN. Technical report, 2018.

M. Ampuja and J. Koivisto. From ‘post-industrial’ to ‘network society’ and beyond: The political conjunctures and current crisis of information society theory. Journal for a Global Sustainable Information Society, 12(2):447–463, 2014.

M. Aschmann, L. Leenen, and J. Van Vuuren. The utilisation of the deep web for military counter terrorist operations. pages 15–22, 2017

V. Braun and V. Clarke. Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2):77–101, 2006. ISSN 14780887. doi: 10.1191/1478088706qp063oa.

U. C. Command. Our history. Retrieved from https://www.cybercom.mil/About/History/.

Darknet Diaries. Ep 50: Operation glowing symphony, 2019. Retrieved from https://darknetdiaries.com/ episode/50/.

R. Deibert, R. Rohozinski, and M. Crete-Nishihata. Cyclones in cyberspace: Information shaping and denial in the 2008 russia–georgia war. Security Dialogue, 43(1):3–24, February 2012.

J. Devanny, L. Goldoni, and B. Medeiros. Strategy in an uncertain domain: Threat and response in cyberspace. Journal of Strategic Security, 15(2):33–47, 2022. doi: 10.5038/1944-0472.

P. Dickinson. The 2008 Russo Georgian war: Putin’s green light, 2021. Retrieved from https://www.atlanticcouncil.org/blogs/ukrainealert/ the-2008-russo-georgian-war-putins-green-light/.

E. Ernst and S. Lee. Countering cyber asymmetry on the korean peninsula: South korea’s defense against cyber attacks from authoritarian sates. Journal for Intelligence, Propaganda and Security Studies, 15(1):165–179, 2021.

Etymonline. Capability, 2023. Retrieved from https://www.etymonline.com/word/capability.

Gov.uk. Joint cyber reserve force, 2023. Retrieved from https://www.gov.uk/government/groups/ joint-cyber-reserve-force.

T. Grant, C. van‘t Wout, and B. van Niekerk. An ontology for cyber ISTAR in offensive cyber operations. In European Conference on Information Warfare and Security, ECCWS, volume 2020-June, pages 117– 125, 2020. ISBN 9781912764617. doi: 10.34190/EWS.20.066.

J. Hunker. ”Cyber war and cyber power - Issues for NATO doctrine”, Reseach Division. NATO Defense College, no. 62, pp. 1–12, 2010. https://www.files.ethz.ch/isn/124343/rp_62.pdf, (2022-03-07).

G. Huskaj. Digital Geopolitics: A Review of the Current State. In International Conference on Cyber Warfare and Security, volume 18, pages 152–161, feb 2023. doi: 10.34190/iccws.18.1.955.

F. Kaplan. ‘wargames’ and cybersecurity’s debt to a hollywood hack. Retrieved from https://www.nytimes.com/2016/02/21/movies/wargames-and-cybersecuritys-debt-to-a-hollywood-hack.html.

F. Kaplan. Dark Territory: The Secret History of Cyber War. Simon Schuster Paperbacks, Harlow, 1 edition, 2017. ISBN 1476763267.

R. Mandel. Optimizing Cyberdeterrence: A Comprehensive Strategy for Preventing Foreign Cyberattacks. Georgetown University Press, 2017.

Microsoft. Defending ukraine: Early lessons from the cyber war. Technical report, Microsoft, June 2022.
Retrieved from https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE50KOK.

Military Wiki. Land information assurance group, 2023. Retrieved from https://military-history.fandom.com/wiki/Land_Information_Assurance_Group.

C. on Foreign Relations. Cyber operations tracker. Retrieved from https://www.cfr.org/cyber-operations/.

PwC and BAE Systems. Operation Cloud Hopper. Technical Report April, 2017. Retrieved from https://www.pwc.co.uk/cyber-security/pdf/cloud-hopper-report-final-v4.pdf.

D. Rohret. Proactive cyber initiative: An expert system framework. In the 9th European Conference on Information Warfare and Security 2010, ECIW 2010, pp. 378–388.

A. Tellis, J. Billy, C. Layne, M. McPherson, and J. Sollinger. Measuring National Power in the PostIndustrial Age. RAND Corporation, first edition, 2000a.

A. J. Tellis, J. M. Sollinger, C. Layne, J. L. Bially, and M. McPherson. Measuring National Power in the Post-Industrial Age. RAND Corporation, first edition, 2000b. ISBN 0833027921.

R. Turn and W. Ware. Privacy and Security in Computer Systems: The vulnerability of computerizedinformation has prompted measures to protect both the rights of individual subjects and the confidentiality of research data bases. American Scientist, 63(2), 1975. Retrieved from https://www.jstor.org/ stable/27845364.

U.S. Cyber Command. U.S. conducts first hunt forward operation in lithuania, 2022. Retrieved from https://www.cybercom.mil/Media/News/Article/3020430/us-conducts-first-hunt-forward-operation-in-lithuania/.

J. van Haaster. ”Assessing cyber power.” In 2016 8th International Conference on Cyber Conflict (CyCon), vol. 61, pp. 7–21, May 2016. doi: 10.1109/CYCON.2016.7529423.

S. Wilkinson. Qualitative research: Theory, method and practice. Sage Publications Ltd., London, 1 edition, 1997.

M. Willett. ”Assessing cyber power”, Survival, vol. 61, no. 1, pp.85–90, 2019.

A State-of-the-art of Scientific Research on Disinformation

Gazmend Huskaj[1,2] and Stefan Axelsson[2]
[1]Geneva Centre for Security Policy, Geneva, Switzerland
[2]Department of Computer and Systems Sciences, Kista, Sweden

g.huskaj@gcsp.ch
stefan.axelsson@dsv.su.se

Abstract: Technological advancements in information and communications technologies and related hardware and software have positively transformed the political, military, economic and social domains in all countries around the globe. These technologies are imperfect, and States and state-sponsored threat actors are exploiting flaws in hardware and software for various types of attacks. Furthermore, the same threat actors exploit software technologies to spread disinformation and disseminate false information to mislead public opinion. This research article reviews the discourse of the scientific community on disinformation. The purpose is to understand where the research focus lies and who the researchers are the co-authors, and the publication venues. This research article reviews the scientific literature using the computational literature review, a semi-automated review method and the structural topical modelling framework to understand trends in the research. Of 3 097 documents published in 1 700 publication venues between 1974 to 2022, 704 were analysed. The results reveal 46 topics on issues such as rumours and disinformation spread during the Covid-19 pandemic, Soviet and Russian Information Warfare, and Trolls and health-related themes and effects.

Keywords: Disinformation; computational literature review; structural topic modelling; social media, trends.

1. Introduction

Technological advancements in information and communications technologies and related hardware and software have positively transformed the political, military, economic and social domains in all countries around the globe (Huskaj, 2023; UNODA, n.d.). These technologies are not without flaws. Threat actors are taking the opportunity to exploit flaws in hardware and software for espionage, attacks, and cybercriminal reasons (Huskaj, 2023). Furthermore, States and state-sponsored threat actors also exploit software technologies like social media to spread disinformation (Landon-Murray, Mujkic & Nussbaum, 2019; Hughes et al., 2021; Wawrzuta et al., 2021). The Great Soviet Encyclopaedia (the BSE) "defines disinformation as the dissemination of false information to mislead public opinion through print and radio" (in Russian: "В БСЭ дезинформация определяется как распространение через печать и радио ложных сведений с целью введения в заблуждение общественного мнения", Litres, 2022).

In the definition, "print and radio" highlight the importance of the role of technology in disinformation. Therefore, understanding the state-of-the-art scientific research on disinformation and the focus of the research is essential for academics, policymakers, and the public. Thus, this review of the scientific literature addresses the research question, *What is the current state of scientific research on disinformation using computational topic models for review and descriptive analysis of scientific articles?*

Although previous research has provided research results into "the psychology of fake news" (Pennycook & Rand, 2021), how various social media platforms are used for disinformation (e.g., Cheung-Blunden et al., 2021; Strudwicke & Grant, 2020), and its impact (e.g., Landon-Murray, Mujkic & Nussbaum, 2019; Schirch, 2021), there is space for a comprehensive computational systematic review that reviews and synthesises the current state of scientific research on disinformation. More specifically, this computational review provides research output per year from 1974 to 2021; the top cited articles in the data set and the top publication venues; a social network analysis of researchers and their co-authors; and through topic modelling, the number of topics where word clouds visualise each topic; and finally, trend topics over time.

Although many review methods, such as a systematic literature review (e.g., Higgins et al. 2019)., a meta-analysis (Chowdhury et al., 2014)., and meta-synthesis (e.g., Noblit et al., 1988) exist, the CLR augments the human researcher and complements "the limitations of human data processing" (Mortenson & Vidgen, 2016, p. 1248).

2. Methods and Materials

The source for collecting research articles on disinformation was Scopus. The Scopus® database is 'the largest abstract and citation database of peer-reviewed literature' (Elsevier, 2018). The nature of Scopus, a database that contains curated research information in a specific format, makes it ideal for downloading and processing the information computationally. On May 31st, 2022, the keyword disinformation was used in the article title,

abstract and keywords. The database presented 3 097 document results from 1974 to 2022, published in 1 700 publication venues. Figure 1 presents published articles per year from 1974 to 2021. The documents were cleaned to remove documents without author, abstract, or doubles.

A first review of the 2 700 document results using the computational literature review (CLR) method revealed that not all results were about disinformation. Instead, some results were about disinformation in information systems, health and palaeontology. The results of the first review of the 2 700 documents using the CLR revealed 78 topics where some of the topics were about States such as America, China, the Czech Republic, Germany, Russia, and the Soviet Union. The topics included organisations such as the KGB and Stasi and "atomic agents" like "Perseus", an alleged KGB spy who had penetrated the American "Manhattan"-project. Furthermore, the topics included propaganda, warfare, influence, and narratives. Thus, the next step was to use these topics as keywords to review what results from the Scopus-database would produce. The keywords used in the title, abstract and keywords were:

```
disinformation AND ( america OR china OR czech OR german OR russia
OR soviet OR kgb OR stasi OR propaganda OR warfare OR influence
OR narrative* )
```

The Scopus-database produced 942 document results. A manual review of the 942 document results revealed three documents being out of scope. Thus, the CLR was applied to the remaining 939 documents, spanning from 1981 to 2022 over 618 publication venues. After removing those without abstract or author, eight hundred ninety-nine (899) documents remained. After a first pass of testing, the results showed that a granular analysis had to be done. All articles from 2004 and onwards were reviewed, but the results still needed to provide more granular differences on how topics were discussed over time. Therefore, all articles (704) between 2015 to 2021 were reviewed. Figure 2 presents articles published per year between 2015-2021.

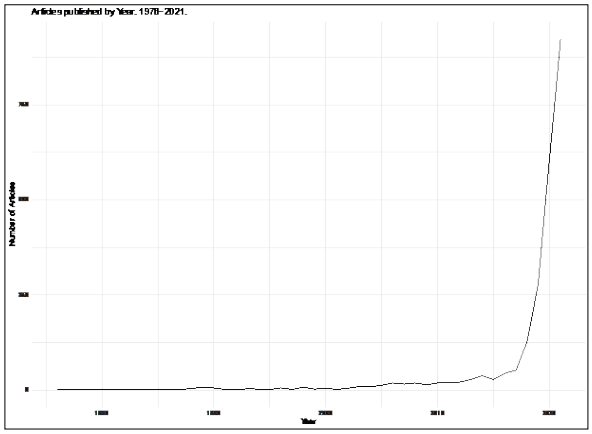


Figure 1: Research articles published per year from 1974 to 2021.

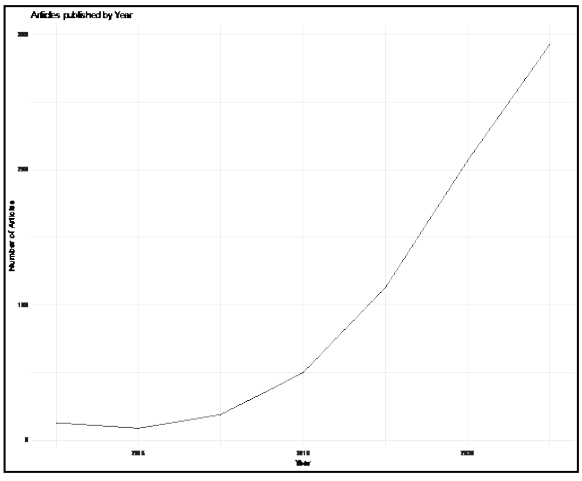


Figure 2: Research articles published per year from 2015 to 2021.

Part of the topic modelling process is to identify the relevant number of topics, k, in the data set. Before designing algorithms to support the identification of k, researchers had to assess a possible k between, for example, 30, 60 and 90. However, the ldatuning-algorithm supports identifying the possible number of topics in a data set. Figure 3 presents the results of applying the ldatuning-algorithm from 10 to 200 topics, by steps of 10. The algorithm output suggests that the data set may have around 50 topics. The next step is applying the ldatuning-algorithm to narrow the search by searching from 30 to 80 in steps of 1. Figure 4 presents that the data set consists of 46 topics.

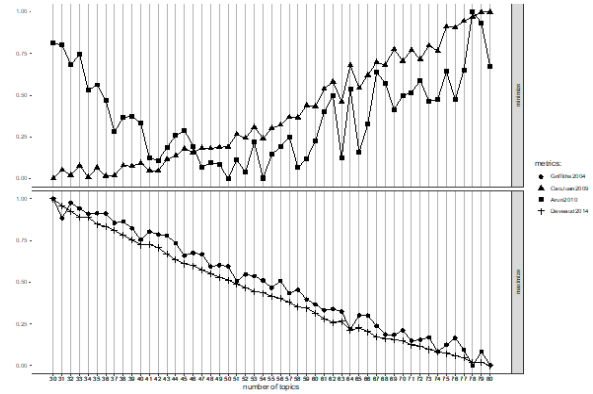


Figure 3: The ldatuning-algorithm to identify the number of topics in the dataset. The x-axis depicts no. of topics, the y-axis depicts algorithm-metrics.

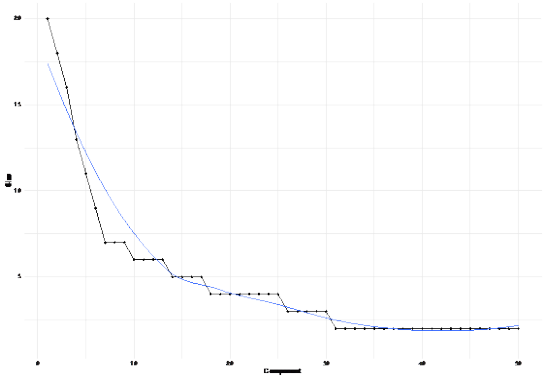


Figure 4: Presenting the 50 first co-author networks, with the top cluster consisting of 20 co-authors.

The results of the ldatuning-algorithm present that 46 topics exist in the dataset. Both LDA and CTM algorithms may be used to present each topic in word clouds. Word clouds are a method to present text visually. They present differences, similarities, and which words are most likely in a topic. The colours show different categories, and the font size presents the likelihood of which words will occur. An example of a word cloud is presented in Figure 4. It is up to the research analyst to infer the word clouds and provide a relevant label for the topic they discuss.

Finally, each topic consists of a set of research documents. These documents are statistically clustered into topics through the application of the LDA or CTM. Therefore, each topic, labelled by the research analyst, can be further analysed to present which research documents exist within that topic. Table 3 presents the insights provided by the 46 topics.

Word clouds are a method to visually present text. They visualise similarities, differences and present which words are most likely to occur in a topic. The colours represent categories while font size represents which words are most likely to occur. Figures 5 and 6 presents two topics that emerged from the data set. The supplementary material provides word clouds of all 46 topics.

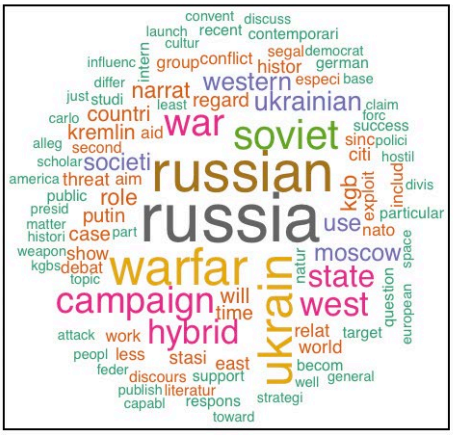


Figure 5: Topic 2 discusses Russian Information Warfare and its Impact on Ukraine and the West.

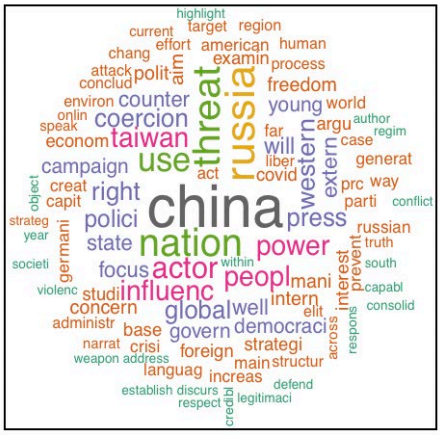


Figure 6: Topic 38 discusses Information Warfare, Disinformation, and External Threats in a Global Context.

3. Results

The results from the computational literature review are publications per year (Figure 1 and Figure 2); most cited articles (Table 1); publication venues (Table 2), and social network analysis of the authors (Figure 6 and Figure 7). Topic modelling provides 46 topics which are evaluated with the word clouds and key terms. The word clouds, topics, and associated papers are presented in the supplementary material. Table 3 presents the results of the qualitative analysis of the 46 topics. Each topic is generated on the documents containing words with the highest probability being linked to that topic. Figure 7 presents trend topics over time.

Table 1: Top 5 articles ranked by citation count.

No.	Authors	Title	Year	Source	Cites
1	Hua J., Shaw R.	Corona virus (Covid-19) "infodemic" and emerging issues through a data lens: The case of China	2020	International Journal of Environmental Research and Public Health	224
2	Ahmad A.R., Murad H.R.	The impact of social media on panic during the COVID-19 pandemic in Iraqi Kurdistan: Online questionnaire study	2020	Journal of Medical Internet Research	218
3	Zellers R., Holtzman A., Rashkin H., Bisk Y., Farhadi A., Roesner F., Choi Y.	Defending against neural fake news	2019	Advances in Neural Information Processing Systems	126
4	Su Z., McDonnell D., Wen J., Kozak M., Abbas J., Šegalo S., Li X., Ahmad J., Cheshmehzangi A., Cai Y., Yang L., Xiang Y.-T.	Mental health consequences of COVID-19 media coverage: the need for effective crisis communication practices	2021	Globalization and Health	101
5	Fallis D.	What is disinformation?	2015	Library Trends	101

Table 2: The Top 5 publication venues ranked by number of articles and citation count.

Rank	Venue	Articles	Rank	Venue	Cites
1	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	22	1	International Journal of Environmental Research and Public Health	254
2	Political Communication	8	2	Political Communication	247
3	Social Media and Society	8	3	Journal of Medical Internet Research	229
4	Media and Communication	8	4	Globalization and Health	132
5	CEUR Workshop Proceedings	8	5	Proceedings of the ACM on Human-Computer Interaction	129

Table 3: The insights generated by the 46 topics identified by the algorithm.

Topic no.	Insights
1. Rumours, Disinformation, and Government Strategies during Crisis and Outbreaks	Rumours and Disinformation during COVID-19 (1.1, 1.2, 1.4, 1.12), Government Strategies and Social Media Use to Combat Disinformation (1.3, 1.5, 1.9, 1.10), Disinformation and Vulnerabilities in Different Geopolitical Contexts (1.6, 1.7, 1.8, 1.11, 1.13).
2. Russian Information Warfare and its Impact on Ukraine and the West	Historical and Cold War Perspectives on Russian Information Warfare (2.1, 2.5, 2.8, 2.11, 2.12, 2.15, 2.18), Russian Propaganda and Influence in Ukraine (2.2, 2.3, 2.10, 2.16, 2.23), Russia's Hybrid Warfare, Cyber Strategy, and Effects on the West (2.4, 2.6, 2.7, 2.9, 2.13, 2.14, 2.17, 2.19, 2.20, 2.21, 2.22).
3. Perceptions, Trust, and Effects of Online Information and Misinformation	Effects of Emotions, Repetition, and Platform Policies on Information Perception (3.4, 3.5, 3.10), Assessing Trust and Influences in Online Information and Alternative Medicine (3.1, 3.3, 3.6, 3.12), Media Practices, Astroturfing, and Access to Truthful Information (3.2, 3.7, 3.8, 3.9, 3.11).
4. Digital Democracy and Disinformation: Impacts on Law, Policy, and Public Perception	Disinformation, Manipulation, and Policy in the Digital Era (4.1, 4.3, 4.5, 4.7, 4.8, 4.9, 4.10, 4.12, 4.14, 4.16, 4.17), Legal Responses and Regulations against Disinformation (4.2, 4.11, 4.15), Ethical, Cultural, and Social Aspects of Digital Information and Public Perception (4.4, 4.6, 4.13).
5. Health Communication, Disinformation, and Public Behaviour in the Digital Age	Health Education, Communication, and Information Seeking Behaviour (5.1, 5.2, 5.3, 5.8, 5.10), Russian Trolls and Disinformation Tactics (5.4, 5.5, 5.6, 5.7, 5.9, 5.13, 5.15, 5.16), Topic Modelling and Analysing Disinformation in Various Contexts (5.11, 5.12, 5.14).
6. Digital Disinformation and its Impact on International Relations and Public Perception	Theoretical Approaches to Digital Disinformation and its Effects (6.1, 6.3, 6.8, 6.11, 6.12, 6.13), Case Studies and Regional Perspectives on Disinformation (6.2, 6.4, 6.9, 6.10, 6.14, 6.17), Digital Media, Public Relations, and Organizational Challenges (6.5, 6.6, 6.7, 6.15, 6.16, 6.18, 6.19).
7. Understanding and Countering the Effects of	User Perceptions and Susceptibility to Misinformation (7.2, 7.4, 7.8, 7.9, 7.12, 7.13), Approaches and Interventions for Countering Misinformation (7.1, 7.3, 7.6, 7.10, 7.11, 7.14, 7.15), Analysing the Effects of Misinformation and Retractions (7.5, 7.7).

Topic no.	Insights
Misinformation and Disinformation	
8. Information Warfare and Cyber Threats in a Technologically Advanced World	Theories, Frameworks, and Techniques in Information Warfare (8.1, 8.2, 8.5, 8.8), Psychological Aspects and Decision-Making in Cyber Environments (8.3, 8.4, 8.6, 8.9), Cybersecurity and Combating Disinformation (8.7, 8.10, 8.11, 8.12, 8.13, 8.14).
9. Disinformation and Communication Strategies in International Contexts	Impact of Disinformation on Global Affairs (9.1, 9.3, 9.5), Case Studies in Culture, Media, and Society (9.2, 9.4, 9.6, 9.7), Countering Disinformation and Fake News (9.8, 9.9, 9.10).
10. Knowledge Management and Disinformation in the Information Age	Effects of Disinformation and Counter-Knowledge (10.1, 10.3, 10.4), Citizen Engagement and Activism in the Digital Era (10.2, 10.5, 10.6), Historical Perspectives and Multidisciplinary Approaches (10.7, 10.8, 10.9), Artificial Intelligence, Journalism, and Combating Misinformation (10.10, 10.11, 10.12).
11. Modelling, Data Analysis, and Deception Detection	Modelling and Data Analysis (11.1, 11.8), Deception Detection and Disinformation (11.2, 11.3, 11.4, 11.6, 11.7, 11.12, 11.14, 11.15), Evaluating and Analysing Historical and Political Data (11.5, 11.10), Understanding Influence (11.9, 11.11, 11.13).
12. Misinformation, Climate Change, Media	Understanding and Analysing Discourses in Sociolinguistics and Communication (12.1, 12.8, 12.15), Disinformation, Misinformation, and Malinformation (12.2, 12.6, 12.9, 12.12, 12.16, 12.17, 12.18), Climate Change Communications, Denial, and Narratives (12.3, 12.4, 12.5, 12.11, 12.13), The Role of Media and Propaganda in Shaping Reality (12.7, 12.10, 12.14, 12.19).
13. Conspiracy Theories and Narratives during COVID-19	COVID-19 Conspiracy Theories (13.1, 13.2, 13.5, 13.10, 13.11, 13.16), Fake News, Disinformation, and the Influence on Public Beliefs and Health Guidance (13.3, 13.7, 13.8, 13.12, 13.14), Conspiracy Theories in Political and Media Contexts (13.6, 13.9, 13.13, 13.15), Conspiracy Theory Research (13.4).
14. Elections, Campaigns, and Social Media	Online Political Advertising, Disclosure, and Regulations (14.1, 14.19), Disinformation, Fake News, and the Influence on Elections and Political Campaigns (14.2, 14.7, 14.10, 14.11, 14.14, 14.16, 14.17), Election Interference, Cyber Propaganda, and Foreign Influence (14.3, 14.5, 14.9, 14.20), Sentiment Analysis and the Use of Social Media in Elections (14.8, 14.13, 14.15, 14.18), Electoral Malpractices, Ethics, and the Role of Journalism (14.4, 14.6, 14.12).
15. Deepfakes, Disinformation, and Their Influence on Individuals and Society	Deepfakes, Artificial Intelligence, and Psychological Security Threats (15.3, 15.8, 15.12), Understanding and Managing Misleading Information and Disinformation (15.1, 15.4, 15.5, 15.6, 15.9, 15.10, 15.11, 15.14, 15.15), Fake News, Deepfakes, and Their Impact on Citizen Engagement and Political Stability (15.7, 15.13, 15.14), Media Narratives (15.2).
16. The Impact of COVID-19 on Vaccine Perception and Public Health	Vaccine Hesitancy and Public Attitudes (16.1, 16.2, 16.11), Public Perceptions and Pandemic Experiences (16.3, 16.4, 16.7), Communication, Misinformation, and Infodemic (16.5, 16.6, 16.13, 16.15, 16.16, 16.19, 16.21, 16.22, 16.23, 16.24, 16.25, 16.26).
17. Dissecting Disinformation and Intent on Twitter	Detection and Analysis of Misbehaviour and Fake News (17.1, 17.4, 17.6, 17.11), Political Manipulation and Extremism through Tweets (17.2, 17.3, 17.5, 17.9, 17.10, 17.14), Disinformation Operations and Emotional Engagement (17.7, 17.12, 17.13), Discursive Characterization and Framing of Narratives on Twitter (17.8).
18. Examining Misinformation and Disinformation in Online Communities	COVID-19 Misinformation and its Global Impact (18.1, 18.3, 18.9, 18.10, 18.15, 18.18, 18.19, 18.24), Detection and Tracking of False Information on Social Media (18.2, 18.5, 18.6, 18.12, 18.21, 18.22, 18.23), The Role of Librarians and Content Moderators in Combating Misinformation (18.8, 18.19), Analysing Disinformation Campaigns and Propaganda in Social Media Networks (18.4, 18.7, 18.11, 18.13, 18.14, 18.17, 18.20).
19. Disinformation and Espionage in Historical Context	Soviet Disinformation and Espionage Cases (19.2, 19.6, 19.8, 19.10), Asbestos Industry and Misrepresentation of Health Risks (19.5), Stasi, AIDS Disinformation, and World War II (19.3, 19.6), Raoul Wallenberg and Diplomatic Controversies (19.6, 19.8, 19.10), International Espionage and Double Agents (19.7, 19.9).
20. Foreign Influence and Disinformation in Elections	Foreign Interference and Election Meddling (20.1, 20.2, 20.3, 20.7, 20.10, 20.11, 20.12), Influence Minimization and Estimation in Social Networks (20.4, 20.5, 20.6), Fake News Detection and Analysis (20.9, 20.13, 20.14), Disinformation Campaigns on Social Media Platforms (20.7, 20.15, 20.16), Case Studies of Disinformation in Elections (20.8, 20.11, 20.14).
21. Techniques and Challenges in Detecting and Classifying Misinformation and Propaganda in Multilingual Content	Multilingual and Multi-ideology Detection Techniques (21.1, 21.5, 21.9), Text Classification and Machine Learning Approaches for Misinformation Detection (21.2, 21.6, 21.7, 21.8, 21.10, 21.14), Challenges in Detecting and Defending Against Neural Fake News (21.3, 21.4, 21.11, 21.12), Identifying State-Funded News and Propaganda (21.13).
22. Addressing Hate Speech, Disinformation, and Online Regulation in the Digital Age	The Impact of Liberalism and Democratic Participation in the Digital Era (22.1, 22.5, 22.11), Legal and Regulatory Approaches to Hate Speech and Disinformation (22.2, 22.7, 22.8, 22.12, 22.13, 22.15, 22.16), Understanding and Countering Online Extremism and Disinformation Tactics (22.3, 22.4, 22.6, 22.9, 22.14), Moderation Practices and Challenges in Newsrooms and Online Platforms (22.10).
23. The Intersection of Peace, Politics, and Disinformation in Modern Society	Economic Growth and the Impact of New Frontiers (23.1), The Role of Social Media in Political Decision-Making (23.2, 23.6), Political Participation, Disinformation, and Public Infrastructure (23.3), Peace Negotiations, Referendums, and Political Opposition (23.4), The Effects of Disinformation on Journalism and Democracy (23.5), Strengthening Resilience in Political Institutions and Processes (23.7).

Topic no.	Insights
24. Understanding Social Bots and Their Impact on Information Dynamics	Detecting and Analysing Social Bots (24.1, 24.2, 24.3, 24.4, 24.10, 24.12, 24.13, 24.17), Influence and Tactics of Bots in Political and Social Contexts (24.6, 24.8, 24.11, 24.14, 24.15, 24.16, 24.18, 24.20, 24.21), Uncovering Coordinated Activities and Disinformation Networks (24.5, 24.7, 24.9, 24.19, 24.22, 24.23),
25. Deception and Manipulation in Political Communication and Propaganda	Euphemism and Deception in Political Discourse (25.1, 25.2, 25.5), Propaganda, Religion, and Literary Intelligence (25.3, 25.4), Manipulation of Information in Historical Contexts (25.6), Politics, Budgets, and Propaganda (25.7).
26. Understanding Psychological and Educational Aspects of Disinformation and Manipulation	Educational Strategies and Media Literacy (26.1, 26.4, 26.7, 26.10), Psychological Mechanisms and Implications of Disinformation, (26.2, 26.5, 26.9, 26.11, 26.12, 26.13), Cybersecurity, Digital Threats, and Countermeasures (26.3, 26.6, 26.8, 26.14).
27. Disinformation and Narrative Manipulation in European Contexts	Disinformation and Narratives in the Czech Republic and Sweden (27.2, 27.5, 27.8, 27.9, 27.15, 27.21, 27.22), Russian Influence, Propaganda, and Disinformation in Europe (27.1, 27.4, 27.10, 27.12, 27.14, 27.16, 27.17), Sociopolitical Narratives, Moral Rhetoric, and Identity (27.3, 27.6, 27.7, 27.11, 27.18, 27.19, 27.20).
28. Public Health, Misinformation, and Vaccine Hesitancy	Vaccine Refusal and Hesitancy (28.1, 28.8), Misinformation and Its Impact on Public Health and Society (28.2, 28.4, 28.5), Healthcare Access and Adaptation in Challenging Contexts (28.3, 28.9), Public Communication and Diplomacy (28.6, 28.7).
29. Russian Influence and Election Interference	Theoretical Perspectives on Interference and Democracy (29.1, 29.5, 29.6), Russian Involvement in Foreign Elections and Democratic Processes (29.3, 29.4, 29.8, 29.9, 29.13), Disinformation Campaigns and Propaganda (29.2, 29.7, 29.10, 29.11, 29.12, 29.14, 29.15, 29.16), Case Studies on Russian Influence and Disinformation (29.17, 29.18).
30. Disinformation and Content Bias on Digital Platforms	YouTube Algorithms and Vaccine Content (30.1, 30.3), YouTubeTracker and Information Operations (30.2, 30.5, 30.6), Bias and Disinformation in Video Recommendations (30.4, 30.7, 30.8), Fake News, Disinformation, and Education (30.9, 30.10, 30.11, 30.12, 30.13).
31. Propaganda, Disinformation, and Education in the Context of Cold War and Modern Conflicts	Cold War Espionage and Disinformation (31.1, 31.6, 31.7, 31.14, 31.16), Digital Literacy, Education, and Democracy (31.4, 31.9, 31.15), Media and Propaganda (31.3, 31.5, 31.8, 31.10, 31.11), Cyber Warfare and Disinformation (31.2, 31.12, 31.13).
32. Detecting and Mitigating Disinformation and Fake News in the Digital Age	Detection Techniques for Fake News and Disinformation (32.1, 32.3, 32.4, 32.6, 32.7, 32.8, 32.9, 32.10, 32.19, 32.21), Analysis, Identification, and Fact-Checking (32.2, 32.5, 32.11, 32.12, 32.13, 32.14, 32.15, 32.18, 32.20, 32.22), Cyber-Enabled Information Warfare and Propaganda (32.17), Assessing Online Information Quality for Specific Domains (32.16).
33. Understanding and Combating Disinformation and Fake News in Online Media	The Role of Trust, Source Credibility, and Motivated Reasoning (33.3, 33.8, 33.9, 33.10, 33.13, 33.14), The Impact of Disinformation on Sociopolitical Dynamics and Protests (33.1, 33.5, 33.7), Media Literacy, Critical Thinking, and Investigative Journalism (33.4, 33.6, 33.11, 33.15), Case Studies of Disinformation Campaigns and Their Effects (33.2, 33.12).
34. Disinformation and Information Warfare in Contemporary Society	Theoretical Approaches to Disinformation and Information Warfare (34.1, 34.3, 34.4, 34.13, 34.14, 34.20, 34.25), The Role of Civil Society and Government Policies in Addressing Disinformation (34.5, 34.7, 34.16, 34.19), Legal, Forensic, and Regulatory Perspectives on Disinformation (34.9, 34.10, 34.12, 34.17, 34.18, 34.21, 34.22), Case Studies on Disinformation, Propaganda, and Information Influence (34.2, 34.6, 34.8, 34.11, 34.15, 34.23, 34.24).
35. Disinformation Campaigns and Platform Use in Information Warfare	Theoretical Approaches and Assessments of Disinformation Campaigns (35.1, 35.4, 35.9, 35.14), Analysing and Tracking Coordination in Disinformation Campaigns (35.2, 35.3, 35.10, 35.12, 35.13), Case Studies of Internet Research Agency (IRA), and Propaganda Tactics (35.5, 35.7, 35.8, 35.11, 35.14, 35.15), Visual Framing and Self-Presentation in Social Media Propaganda (35.6).
36. Students, Information Literacy, and Combating Misinformation in the Digital Age	Information Literacy and Education (36.4, 36.8, 36.11, 36.12, 36.13, 36.15), Online Information Seeking, Sharing, and Consumption Habits (36.1, 36.3, 36.6, 36.7, 36.10, 36.16), Understanding and Tackling Misinformation and Fake News (36.2, 36.5, 36.9, 36.14), Information Disorders and Media Landscape during Crisis (36.5, 36.12, 36.16).
37. Digital Communication, Disinformation, and Critical Approaches in the Information Age	Digital Media, Information Literacy, and Critical Thinking (37.2, 37.6, 37.7, 37.9, 37.14), Disinformation, Information Warfare, and Cognitive Security (37.3, 37.4, 37.11, 37.12, 37.13, 37.17), Political Communication, Manipulation, and Branding in the Digital Age (37.5, 37.15, 37.18, 37.19), Comparative Studies on Disinformation Regulation and Resilience (37.1, 37.8, 37.10, 37.16).
38. Information Warfare, Disinformation, and External Threats in a Global Context	China's Cyber Coercion and Disinformation Tactics (38.1, 38.4, 38.5, 38.14), Russian Influence and Information Warfare Strategies (38.3, 38.8, 38.13), Global Dynamics of Disinformation and National Security Implications (38.2, 38.6, 38.9, 38.10, 38.11, 38.12, 38.15, 38.16, 38.17), Social Implications of Information Warfare and Defence (38.7).
39. Fact-Checking, Disinformation, and Combating Misinformation in the Spanish-speaking World and Beyond	Fact-Checking and Disinformation in Spain and Latin America (39.1, 39.2, 39.3, 39.4, 39.5, 39.8, 39.9, 39.11, 39.14, 39.15, 39.19), Social Media, Disinformation, and Fact-Checking Platforms (39.6, 39.10, 39.16, 39.17, 39.18, 39.21), Tools and Techniques for Monitoring and Combating Disinformation (39.7, 39.12, 39.13, 39.20).
40. Truth, Influence, and Disinformation in Politics and Society	Disinformation, Fake News, and Political Implications (40.2, 40.3, 40.9, 40.10, 40.11, 40.12, 40.14), Beliefs, Trust, and Values in Information Evaluation (40.5, 40.13, 40.15), Historical and Cultural Perspectives on Truth and Influence (40.1, 40.4, 40.7, 40.8).

Topic no.	Insights
41. Historical Narratives, Politics, and Media in American and Global Contexts	Historical Events and Narratives in Games and Literature (41.1, 41.5), Political Climate and Disinformation in Contemporary Societies (41.2, 41.4), Interpretative Challenges and Historical Accounts (41.3, 41.6).
42. Disinformation and Narratives in Mainstream Media and Digital Platforms	Disinformation and Opinion Leaders in Digital Media (42.2, 42.4, 42.5, 42.7), State, Society, and Disinformation Warfare (42.3, 42.6), Mainstream Media and Disinformation Analysis (42.1, 42.8, 42.9).
43. Russian Influence and Disinformation across History and Contexts	Russian Disinformation and Propaganda in Historical Context (43.1, 43.7, 43.8, 43.9, 43.12), Disinformation and Misinformation in Contemporary Society (43.2, 43.13, 43.15, 43.16, 43.17), Weaponization of Information and Targeting Specific Groups (43.3, 43.4, 43.5, 43.6, 43.10, 43.11, 43.14).
44. COVID-19 Pandemic and Disinformation Global Impact and European Perspectives	Origins and Impact of COVID-19 Disinformation (44.1, 44.5, 44.9, 44.10, 44.12), European Union and the Fight against COVID-19 Disinformation (44.7, 44.11, 44.14), Disinformation Beyond the COVID-19 Pandemic (44.2, 44.3, 44.4, 44.6, 44.8, 44.13).
45. Exploring Disinformation in the Digital Age Politics, Academia, and Online Practices	Political Aspects of Disinformation (45.1, 45.3, 45.7, 45.8, 45.9), Online Fake News and Content Analysis (45.2, 45.4, 45.6), Economic Disinformation and Recommendations (45.5).
46. The Impact of Mis- and Disinformation on Journalism	Journalists' Role Perceptions and Disinformation (46.1).

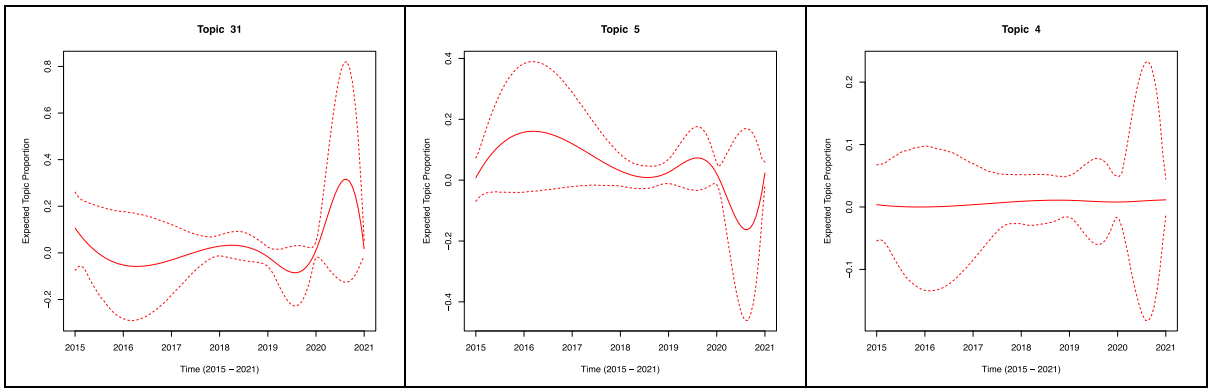


Figure 5: Trends on three topics from 2015 to 2022. Topic 31 – Propaganda, Disinformation, and Education in the Context of Cold War and Modern Conflicts; Topic 5 – Health Communication, Disinformation, and Public Behaviour in the Digital Age; Topic 4 – Digital Democracy and Disinformation: Impacts on Law, Policy, and Public Perception.

4. Discussion

The purpose of this review article was to answer the question: *What is the current state on disinformation using computational topic models for review and descriptive analysis of scientific articles?* The computational literature review and the structured topic modelling show that even though disinformation has been a topic of interest for researchers, it is first from 2016 and onwards that researchers have started to show interest on the topic. Historically, intelligence agencies such as the KGB, and STASI, have been used to conduct disinformation operations on topics like AIDS. Furthermore, researchers have shown interest in intelligence agencies and their operators' activities in countries like Sweden, and allegedly targeting individuals such as Raoul Wallenberg.

Table 4: The domain clusters in descending order.

Domain	Topics	Domain (cont.)	Topics (cont.)
Political	7	History	3
Health	6	Combating	2
Platforms	6	Psychology & Bias	2
Context	5	Targets	2
Fields	4	AI	1
Detection	3	Impact	1
Effects	3	Source	1

Disinformation operations today are different mainly because of cyberspace. Cyberspace, the fifth domain, has increased how people generate information, do business and share and consumes information. The average citizen can access more information than states could 40 years ago. Furthermore, information about people, organisations, products, and capabilities, is readily available and pushed online rather than being pulled like 40 years ago. Finally, content based on visualised information (images, film) is much more accessible than 40 years ago. All of this is thanks to the rapid development of information- and telecommunications technologies (ICT),

increased computational power at low cost, low-cost high-storage speeds, and increased hardware and software achievements, especially within machine learning and artificial intelligence that can quickly manipulate images, film and voice: two of five senses.

So, what does all this mean? It means that as humankind is increasing interconnectedness with cyberspace through means such as virtual reality, augmented reality, and mixed reality, threat actors can exploit these new technologies and employ computational means to manipulate information spaces that align with threat actors' intent and, in turn, manipulate what humankind sees and hears. Threat actors can employ multiple avenues of approaches to achieve this goal: Exploit vulnerabilities in cyberspace; use ML/AI-enabled tools such as "deep-fakes" to impact real-world perception; state-sponsored news sites; "research"; buy stakes in companies or buy companies wholly; exploit vulnerabilities in the legal system (lawfare); create dissent between various ethnic groups; exploit religion; diseases and pandemics; use insiders, and "useful idiots". In short, manipulate information to suit the strategic security policy goals and disseminate it through multiple channels.

5. References

Cheung-Blunden V., Sonar K.U., Zhou E.A., Tan C. (2021). Foreign disinformation operation's affective engagement: Valence versus discrete emotions as drivers of tweet popularity. Analyses of Social Issues and Public Policy.

Chowdhury, R., Warnakula, S., Kunutsor, S., Crowe, F., Ward, H. A., Johnson, L., Franco, O. H., Butterworth, A. S., Forouhi, N. G., Thompson, S. G., Khaw, K.-T., Mozaffarian, D., Danesh, J., & Di Angelantonio, E. (2014). Association of dietary, circulating, and supplement fatty acids with coronary risk: A systematic review and meta-analysis. Annals of Internal Medicine, 160(6), 398-406.

Huskaj, G. (2023). Digital Geopolitics: A Review of the Current State.

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions. 2nd Edition. Chichester (UK): John Wiley & Sons, 2019.

Landon-Murray M., Mujkic E., Nussbaum B. (2019). Disinformation in Contemporary U.S. Foreign Policy: Impacts and Ethics in an Era of Fake News, Social Media, and Artificial Intelligence. Public Integrity.

Litres. (2022). Новая ложь взамен старой. Retrieved from: https://www.litres.ru/anatoliy-mihaylovich-golicyn/novaya-lozh-vzamen-staroy/chitat-onlayn/.

Mortenson, M. J., & Vidgen, R. (2016). A computational literature review of the technology acceptance model. International Journal of Information Management, 36(6), 1248–1259. https://doi.org/10.1016/j.ijinfomgt.2016.07.007.

Noblit, G. W., & Hare, R. D. (1988). Meta-ethnography: Synthesizing qualitative studies. Sage Publications.

Schirch L. (2021). Digital information, conflict and democracy. Social Media Impacts on Conflict and Democracy.

Strudwicke I.J., Grant W.J.(2020). #JunkScience: Investigating pseudoscience disinformation in the Russian Internet Research Agency tweets. Public Understanding of Science.

UNODA (n.d.). Developments in the field of information and telecommunications in the context of international security. Retrieved from: https://www.un.org/disarmament/ict-security/.

Wawrzuta D., Jaworski M., Gotlib J., Panczyk M. (2021). Characteristics of antivaccine messages on social media: Systematic review. Journal of Medical Internet Research.

Fake news as a Distortion of Media Reality: Tell-truth Strategy in the post-truth era

Anastasiia Iufereva
IULM University, Milan, Italy

anastasiia.iufereva@studenti.iulm.it

Abstract: The article deals with fake news which has been considered one of the greatest threats to information security. The expansion of digital technologies and the development of communication networks have contributed to the spreading of misinformation. In particular, the emergence of different sources of information on the Internet, the growing polarization of opinions in the political and socio-economic dimensions, the devaluation of the fact, and the widespread fake news on the Internet (e.g., social media) form the question of revision of the process of collecting, verifying presenting information, methods, and technologies for verifying facts, including methods for countering fake news. Although this issue has been widely investigated in academic discourse, there are still controversial arguments regarding which elements should form a tell-truth strategy. This paper focuses on recent research that reflects trends and patterns in this field and on the author's empirical survey - interviews with university professors and media experts (N=6), journalists (N=6), and students (N=14) in Russia. In this study, the author describes the key characteristics of fake news and the elements of this tell-truth strategy. It is intended that this paper focuses on both professional journalists and professors who may use the results of this investigation in such courses as political science, sociology, philosophy, and journalism.

Keywords: fake news, Internet, journalism, social media, truth, post-truth era.

1. Introduction

The essential transformations that have taken place in the mass media sphere over the past decade are due to the diffusion of the Internet and widespread social practices online. This has caused a number of alterations in the journalists' professional culture. In particular, the accustomed algorithms for gathering and presenting news have been modified because of appearing to new actors in the media environment - the people who began to exercise the ability to express themselves online, therefore, boundaries between journalists and their audiences have blurred. As far as traditional dichotomies between message senders and receivers are blended (Marshall et al., 2003), almost everyone who has the access to the Internet can disseminate information creating multiple channels of information and opportunity to present alternative views regarding a certain issue. In these conditions, a ''post-truth era" has been formed that is the crossing at times conflicting perspectives, perceptions, and approaches. The post-truth phenomenon reflects the transformation of public discourse that deals with the distortion of media reality. It causes problems and obstacles that prevent the media from exercising their professional activity in the digital space because truth, facts, and analytical conclusions have less persuasive power than ready-made interpretations. Post-truth is a consequence of occurring postmodern processes (in particular, mediatization), technological transformations (the widespread use of digital means of communication), changes in social communication (competition of influential actors in the information space), and a general decline in trust in social institutions. The destructive consequences of post-truth (in particular, spreading fake news, devaluation of facts, etc.), which destroy public discourse in the media environment, and exacerbate political and social differences, and the resonance that arises around these consequences speak to the growing need for relevant research. Therefore, the main aim of this investigation is to describe specific features of fake news and identify which elements constitute the truth-telling strategies based on the results of theoretical and empirical studies.

2. Literature review

The term "post-truth world" was first introduced by the American publicist S. Tesich in his essay "The Government of Lies" (The Nation, 1992). The attention was paid to not the dishonesty of politicians but the indifferent attitude of voters that try to "escape from the truth". Citizens "didn't want bad news anymore, no matter how true or vital to our health as a nation" (Grachev, Evstifyev, 2020). The term "post-truth" appeared in political discourse later when in December 2011 the economist P. Krugman used this term to describe the statements of the Republican candidate M. Romney, discredited the incumbent President of the USA B. Obama (Krugman, 2011). In 2016, during D. Trump's campaign, the term "post-truth" was widespread because it reflected the political situation in America at that time. Today this term has been widely discussed in academic research because the Internet makes a great contribution to the existence and dissemination of the post-truth. Digital technologies and the destruction of the vertical structure of knowledge production facilitate a more

flattened structure with multiple nodes of information and expression (Waisbord, 2018, p. 5). Digital technologies have given the users of the Internet access to the ability to create and distribute information (Martin, 2017; Pak, 2017). Hence, many actors are involved in the content creation process. The production of a massive flow of information leads to the fact that objective facts are less influential in shaping public opinion than appeals to emotion and personal belief (Sales, 2017). In the context of the post-truth growing number of fake news makes to revise the value of information, facts, and truth.

Fake news is the pivotal topic of relevant research on communication and mass media (Loveless, 2020; Waisbord, 2018; Fitzpatrick, 2018). In academic discourse, scholars face the acute problem of defining fake news. The heterogeneity of the term “fake news” leads to it becoming “a catch-all term with multiple definitions” (Lilleker 2017, p. 1), hence, different approaches to defining specific characteristics of fake news as a phenomenon. For instance, Gelfert (2018) suggests considering a definition of the term “fake news” depending on the context which determines the content of fake news and its pivotal features. Within the first direction, the digital environment (in particular, social media) has been conducive to the creation and distribution of fake news. “Fake news” may be understood as “the online publication” of false statements of fact (Klein et al., 2017, p. 6). Furthermore, a “core feature of contemporary fake news is that it is widely circulated online” (Bakir et al., 2017, p. 1). Some definitions underline the role of the medium, which is a platform by which fake news is distributed: fake news is “authentic material used in the wrong context, news products disseminated on websites that specialize in false news and use layouts similar to those of real media sites and all false information and contents intended to manipulate public opinion, journalism deception as an act of communicating messages either verbally in the form of a lie or non-verbally by withholding information in order to initiate or uphold a false belief” (Sutu, 2019, p. 83). Second, fake news doesn't really have a whole lot to do with the situations that occur in the world. Fake news consists of stories “invented entirely from thin air”, to be “completely fabricated”, to transmit “new content [that] is 100% false”, and to have “no factual basis” (Gelfert, 2018, p. 96). The third direction makes focuses on the purpose of fake news, which is to mislead readers, distort reality, and manipulate public opinion. “An element of deliberateness is imputed to the creation and circulation of fake news, which in turn is deemed “deliberately misleading” and involving “intentionally or knowingly false statements of fact”, “deliberate spread of misinformation”, along with the “intention to deceive”” (Gelfert, 2018, p. 96). Recent studies also identified a few crucial intentions that motivate the creation and dissemination of fake news: financial, ideological, and entertainment. The first one refers to the production of fake news with the intention to earn money, converting clicks on the Internet into advertising revenues (Tandoc, 2019). The ideological aspect is manifested in situations when creators try to discredit certain political actors or states (Allcott, 2017). Entertainment can be noticed in the situation when a product does not have a certain aim as earning money or defaming some actors in the social or political sphere because the motivation is the desire to make a joke to the audience.

Particular attention is paid to the social and political preconditions for the emergence of fake news, the methods, and procedures for creating fake news, and the threats posed by fake news: the spread of the phenomenon of blogs and UGC (Hellmueller, 2013), and increasing competition, especially from social media, which have become a powerful competitor for journalists (Hannan, 2018), the threat of manipulation (Fitzpatrick, 2018) and competition because competition is a factor in driving journalists to "break" a story before anyone else can do this (Fitzpatrick, 2018, p. 10). As a result, many scholars point to a paradigm shift that actualizes the question of articulation of a new occupational norm plays an important role in establishing it (Elliot, 2008; Hellmueller, 2013).

There are currently different approaches to defining truth-finding strategies in the academic literature. For instance, Karlsson draws attention to two different strategies: “the traditional strategy, <…> in which accurate information is conveyed to the audience; and the newer transparency strategy, where truth-telling is created through forthrightness and discourse and is subject to change over time. The first relies on journalists’ scrutiny and the second on openness and user participation” (Karlsson, 2011, p. 284). Other academic research suggests other journalistic normative guidelines that could form the basis of truth-finding strategies: transparency (Allen, 2008; Kovach et al., 2001), objectivity (Hellmüller, 2013), interactivity (Lasica, 2004); accuracy (Fitzpatrick, 2018, p.5), openness and accountability (Karlsson, 2011). At the same time, scholars emphasize that “empirical research now is needed that can systematically explore the many ways in which that is happening” (Singer, 2007, p. 91). This is a fair point because so far there is a small number of studies (Widholm et al, 2019; Hellmueller, 2013) that summarize empirical developments on the topic of truth-telling strategies that have been involved by journalists in their professional activity. Moreover, in these papers also some of the elements of tell-truth strategies have been considered. For instance, it is necessary to pay attention to Widholm’s research (Widholm

et al, 2019) which focuses on objectivity and transparency only but considering these aspects throughout the gender problem that is also important. Other elements of tell-truth strategies are beyond the author's attention. Therefore, there is also a lack of empirical research that reveals the content of methods and algorithms that are used today by professional journalists and ordinary users of the Internet to verify information, as well as which elements should constitute the tell-truth strategies. It is these gaps that the present study aims to fill.

3. Method

The author conducted a survey (November 2022, Yekaterinburg, Russia) - the interview with university professors and experts in the sphere of mass media, communication, sociology, and political science (N=6), professional journalists who work in news outlets (N=6), and students (specializations - journalism; sociological and marketing research) (N=14). Males are three respondents; females are 26 respondents. Age of respondents: from 15 to 20 years old (10 respondents), from 21 to 30 years old (8 respondents), from 31 to 40 (6 respondents), and from 41 to 50 years old (2 respondents). The sample was random. The specialization and the sphere of an occupation of a respondent were considered first. Two research questions were posed: RQ1: How do you understand the term "fake news"? RQ2: In your opinion, what elements (apart from, objectivity and transparency) should constitute the tell-truth-telling strategies? The survey was organized online and conducted using digital platforms (Facebook, and What's app).

4. Results

Based on the results of our research, it was advisable to distinguish three main groups that show the opinions of (a) professors and experts; (b) professional journalists; (c) students.

4.1 Fake news as a concept

(a) According to the opinions of professors and experts in the sphere of mass media and communication, it is noticed that there are two essential lines that are aligned with the core directions presented in our analysis of academic research. On the one side, respondents underline that fake news is a complete distortion of reality where information, facts, and opinions are fabricated in order to mislead the audience and affect public sentiment.

"Fake news is false information, not supported by facts, evidence, photos, or videos. This is a hoax. Often, fake news is planned, deliberately distorting information for certain purposes of a certain social or public group" (female, 47, associate professor, Russia)

*"Fake news is an informational message that can have a strong impact on society's moods and is based on information that cannot be unambiguously verified" (*male, 35 years old, Ph.D. in sociology, independent researcher, Russia)

Respondents admit that apart from fake news that is entirely invented and has no connection with real facts and events, there is news that may be based on real facts, opinions, and events. In this case, fake news may be a partial distortion of reality that can reflect real situations in the world and at the same time include other misinformation.

"Fake news is news that partially misrepresents the real state of affairs or constructs an alternative, completely simulated reality in order to form among the general public or certain target social groups a worldview and behaviour beneficial to certain financial, industrial, political, and other interest groups" (female, 37 years old, Ph.D. in political science, Russia)

"Fake news is news or reports of past events that contain a false interpretation in order to obtain the necessary publicity in order to influence political decisions or public opinion" (male, 47 years old, director of the autonomous non-profit organization "Centre for European and Asian Studies", senior lecturer, Russia)

b) Professional journalists emphasize two different types of fake news. On the part, fake news is information that may be fabricated without references to the real facts, opinions, and facts in order to cause resonance in society. Conversely, it may consist of real facts on which misinformation can be constructed.

"I understand the concept of "fake news" as a deliberately false message about an event or character that did not exist. That is, for me, "fake news" is still more about something that in reality did not exist or did not happen at all. There is a second version of fake news when the event took place. It was "overgrown" with non-existent details that were given out in the news feed as having happened" (female, 31 years old, journalist)

Respondents also point out that the process of distortion of fake news may be unintended when an author does not have the aim of misleading the audience.

"Fake news may be distorted in a natural way (for example, someone misunderstood something, someone paraphrased something, etc.)" (female, 23 years old, journalist)

"<Fake news> may simply be unverified information gained from the source, while the source itself could be mistaken, and in this case, it turns out to be a fake" (male, 25, journalist)

In terms of another view that is along with the normative approach. The main accent is made on ways of creating such news when the basic principles of journalism are violated or neglected.

"Fake news is publications that violate the basic journalistic principles as the publication of different opinions in equal proportions, the absence of an imposed assessment on the part of the author, and social responsibility" (male, 31 years old, journalist)

c) Based on the answers of the students surveyed it was reasonable to identify the following groups in which various features of fake news are underlined.

Fake news is information that comprises non-existing facts and it does not have a link to the world or information that can include content (e.g., facts, opinions, and events) that may be partly true. Besides, some respondents noted that the distribution of such fake news can have unintended effects.

"Fake news is information that is not true" (female, 19 years old, student, specialization - journalism)

"Fake news is information that is completely or partially untrue" (female, 19 years old, student, specialization - journalism)

"Fake news is information that is not true. This can be both unintended and deliberate dissemination of fake news" (female, 19 years old, student, specialization - journalism)

According to another view, one of the essential components of fake news is appealing to emotions instead of facts.

"Fake news is a circumstance in which objective facts are less important in shaping public opinion than appeals to emotions and personal beliefs" (female, 19 years old, student, specialization - journalism)

The next crucial and integral component of fake news is vague sources of information.

"Fake news usually has a very vague source. It seems to me that this is one of the main characteristics" (female, 24 years old, student, specialization - sociological and marketing research)

"In fake news, the author does not indicate the source from which the information was taken, does not provide expert comments" (female, 19 years old, student, specialization - journalism)

4.2 Elements of the tell-truth strategy

a) The significant number of professors and experts surveyed underline that objectivity and transparency are indeed the most essential aspects that should be considered in the work of professional journalists. Submission of different views on the same issue is a guarantee of a more accurate reflection of situations, their consequences, and their subsequent impact on certain aspects of life. Appealing to the different sources is also one of the crucial steps in the preparation of a text. More specifically, it is necessary to pay attention to the wide set of original information (e.g., official reports and statements, links to online list archives, and statistics) that can help avoid misinformation. For instance, one of the experts pointed out that nowadays one of the main sources of fake news is social media (e.g., Telegram channels) where it is complicated to identify if the news is true or not.

"...it is important for experts to treat such messages <on telegram> with a certain degree of skepticism. The corporate standards of many media require the presence of the points of view of all participants in the conflict. However, the main way is still the creation of legal mechanisms for mandatory verification of information.

Additionally, it is still worth starting a discussion about the regulation of the information flow in telegram channels and social networks" (male, 47 years old, director of the autonomous non-profit organization "Center for European and Asian Studies", senior lecturer at the university, Russia)

Apart from these elements, the respondents pointed out other ones that may be a basis for the tell-truth strategy.

"The competent selection of information sources (both documentary and human) by a journalist. Second, the involvement of unbiased experts, if possible (if still such a phenomenon still exists). Third, the journalist's own competence in the issues covered. Finally, the building of an effective mechanism for checking the work of journalists within the editorial office" (female, 37 years old, Ph.D. in political science, Russia).

"I would add here only one more element - the designation by the author of the material of his own position on the issue, which can be taken out in a separate section. Obviously, the personal beliefs of the author of the material can greatly influence how he prepares this material and what conclusions he comes to. Reading such materials, the reader should make an amendment to this" (male, 35 years old, Ph.D. in sociology, independent researcher, Russia)

b) Journalists surveyed also shared the view that the presence of multiple points of view (objectivity) and pointing of source information (transparency) are core elements of tell-truth strategies. For example, some respondents decided to reveal this issue by appealing to the professional aspects of their activity such as technologies of work with information considering these elements.

"You can double-check the information in a higher authority. If, for example, it is said that evidence is being collected on some fact that will be submitted to the international court, then you can go to the website of the court of human rights and see if this is true. All references to alleged documents are also easily verified by official authorities" (female, 31 years old, journalist)

"Presenting multiple points of view is reasonable. If there was any conflict, it is important to reflect on all points of view and request comments from both or all parties. The main thing is to remember that refusal to comment is also a point of view" (female, 22 years old, journalist)

However, along with objectivity and transparency respondents specified the references to other crucial aspects such as numbers, quotes, the principles of openness, and non-value.

"In addition to transparency and objectivity, I would also call reliance on facts (numbers, quotes, examples), otherwise the text looks empty and does not inspire confidence" (female, 23 years old, journalist)

"Valuelessness (if not an author's column), social responsibility, understanding the impact of the text on society, the ethical aspect" (female, 25 years old, journalist).

c) Drawing on the results of the survey with students, it is noteworthy that they concentrate mainly on clarifying features of such elements as objectivity and transparency (additional elements of tell-truth strategies were not suggested in their answers). Respondents describe technologies for obtaining information, paying attention to two specific aspects - the references to the factual material and attracting competent experts.

"It seems to me that it is worth adding more reinforcement of information with photos, videos, and voice recordings of conversations. This will not only attract an audience but will act as an additional "airbag" for the journalists themselves" (female, 19 years old, student, specialization - journalism)

"When preparing journalistic material, it is necessary to study to the smallest detail the problem that needs to be reflected in the publication, turn to experts and get their opinion, find heroes who understand the problem and take their comments. It is also necessary to consider how competent the speaker is in a particular area, that is, to conduct a pre-interview" (female, 19 years old, student, specialization - journalism)

"Sources of information can always be bought by someone, so I think it is necessary, in addition to testimonies, to consult with professionals in this field" (female, 19 years old, student, specialization - journalism)

5. Conclusion

This paper focuses on the results of the analysis of academic literature regarding the following issues: fake news as a concept and the main elements of the tell-truth strategy. It was determined that there are different contexts in which fake news may be defined (e.g., technological, social, or cultural conditions). Then, there are various

approaches to defining the essence of fake news. It may be identified as completely fabricated information or as partially invented content where the real facts, situations, and events are preserved but there is also misinformation that leads to the distortion of the presented reality in the news. Furthermore, the scientific paper is devoted to the elements of the tell-truth strategy: transparency, objectivity, interactivity; accuracy, openness, and accountability.

The author conducted an empirical survey - interviews with Russian university professors and media experts, journalists, and students to describe the key characteristics of fake news and determine elements of this tell-truth strategy. Regarding the first question, it is revealed that the respondents understand the phenomenon of fake news as information that is completely or partially fabricated. At the same time, respondents surveyed underline that the distribution of fake news does not always have intentional character because sometimes it may have an unintentional effect. As for the elements of tell-truth strategies are concerned, the respondents notice the following ones: the competent selection of information sources; the involvement of unbiased experts; the building of an effective mechanism for checking the work of journalists within the editorial office; the scepticism regarding unknown source; paying attention to the references to the factual material (e.g., statistics, foto- and video-materials).

References

Allcott, H. Gentzkow, M. (2017) "Social media and fake news in the 2016 election", *Journal of Economic Perspectives*, 31(2), pp. 211–236.

Allen, D. (2008) "The trouble with transparency", *Journalism Studies*, N0 9(3), pp. 323–340.

Bakir, V., McStay A. (2017), "Fake news and the economy of emotions: problems, causes, solutions", *Digital Journalism*, N 6(2), pp. 1–22.

Elliot, D. (2008) *Essential Shared Values and 21st Century Journalism.* In: L. Wilkins, C. G. Christians (Eds). The Handbook of Mass Media Ethics. New York: Routledge.

Fitzpatrick, N. (2018) "Media Manipulation 2.0: The Impact of Social Media on News, Competition, and Accuracy", *Athens Journal of Mass Media and Communications,* Vol 4, No 1, pp. 45–62.

Gelfert, A. (2019) "Fake News: A definition", *Special Issue: Reason and Rhetoric in the Time of Alternative Facts,* Vol 38, No 1, p. 84–117.

Grachev, M. N., Evstifyev, R. V. (2003) "The Concept of "Destruction of Truth" in the Context of a Digital Society (Analytical Review)", *The Contours of Global Transformations: Politics, Economics, Law,* Vol 13, No. 2, pp. 229–248.

Hannan, J. (2018) "Trolling Ourselves to Death? Social Media And Post-truth Politics", *European Journal of Communication,* Vol 33. N0 2, pp. 214–226.

Hellmueller, L., Vos, T. P., Poepsel, M. A. (2013) "Shifting Journalism Capital?", *Journalism Studies*, Vol 14. No 3, pp. 287–304.

Karlsson M. (2011) "The Immediacy of Online News, The Visibility Of Journalistic Processes and a Restructuring of Journalistic Authority", *Journalism: Theory, Practice & Criticism*, Vol 12, No. 3, pp. 279–295.

Klein, D., Joshua R. W. (2017) "Fake news: a legal perspective", *Journal of Internet Law,* 20(10), pp. 5–13.

Kovach, B., Rosenstiel, T. (2010) *How to Know What's True in the Age of Information Overload.* Publisher: Bloomsbury.

Krugman, P. (2011) "The Post-Truth Campaign", [online], The New York Times, https://www.nytimes.com/2011/12/23/opinion/krugman-the-post-truth-campaign.html

Lilleker, Darren G. (2017) "Evidence to the Culture, Media and Sport Committee 'Fake news' inquiry presented by the Faculty for Media & Communication, Bournemouth University. Documentation. Bournemouth University [online], http://eprints.bournemouth.ac.uk/28610/

Loveless, M. (2020) *Information and Democracy: Fake News as Emotional Weapon*, in: Democracy and Fake News: Information Manipulation and Post-Truth Politics", London.

Marshall, P., Burnett, R. (2003) *Web Theory: An Introduction.* Publisher: Routledge, London.

Martin, N. (2017) "Journalism, the pressures of verification and notions of post-truth in civil society", *Cosmopolitan Civil Societies: An Interdisciplinary Journal,* 9(2), 26, pp. 41–55.

Pak, C. (2017) "News companies link sharing on Twitter as informative advertising and content signaling", *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, Denver, Colorado, pp. 312–315.

Sales, L. (2017) "New ACTU boss says it's OK for workers to break "unjust laws" [online], Report, Sydney, http://www.abc.net.au/7.30/content/2017/s4636757.htm

Singer, L. M., Alexander, P. A. (2017). "Reading on Paper and Digitally: What the Past Decades of Empirical Research Reveal". *Review of Educational Research,* No 87(6), pp. 1007–1041.

Sutu, R. (2019) "Fake News, from Social Media to Television Case Study of the Romanian Presidential Elections 2019", *Styles of Communication,* Vol 11, N0 2, pp. 81–92.

Tandoc, E.C. (2019) "The facts of fake news: A research review", *Sociology Compass,* pp.1-9.

Waisbord, S. (2018) "Truth is What Happens to News", *Journalism Studies,* Vol 19, No 13, pp. 1866–1878.

Hybrid Threats-Possible Consequences in Societal Contexts

Georgiana-Daniela Lupulescu
National Defense University "Carol I", Bucharest, Romania

geo.lupulescu@yahoo.co.uk

Abstract: Hybrid threats have become a persistent term in the 21st century geopolitical architecture, acquiring new values as innovative unconventional means come to be used by both state and non-state actors in contemporary conflicts, with a view to obtaining strategic advantages, yet with devastating consequences at individual level. While the armed conflict effects have long been studied, the war metamorphosis with hybrid threats innuendos bring new challenges in assessing societal consequences, even more so, as they are increasingly identified in apparently peaceful times. A multifaced perspective on the threat outcome reveals multiple latent consequences, such as physical, material, psychological and emotional ones. Fear, one of the dominant human emotions, is the first to be triggered when any threat is present, regardless of its occurrence probability or possible effects. Fear becomes a strong drive for individuals' future actions, sometimes prompting an offensive or defensive reaction previously embedded in the main actor's behavior. In this context, the present paper aims to identify, analyze and understand the Russian-Ukrainian conflict consequences on the European states' neighboring population, looking at the reactions and decisions triggered by fear. Using observation as a research method but also the case study method, we identified a series of similarities and differences in these countries' reaction to solving situations, migration- generated crises, Russian disinformation and propaganda and Ukraine or other European state oriented cyber-attack. The main goal for this approach is to highlight the hybrid threats emotional consequences in conflicts that are more than psychological. Moreover, this is a preliminary step in a PhD research thesis with a view to provide states with solutions for resilience policies, to ensure their citizens' survival and well-being.

Keywords: Hybrid Threat, Russia, Psychological Warfare, Disinformation, Cyber-Attacks

1. Introduction

The hybrid term appeared as a necessity to define another kind of war, one that includes all that we previously knew about war and peace, but also the development of technology. Although there is not a widely accepted definition, scholars, policy makers and practitioners use the term to describe a combination of military actions and non-conventional tactics, using proxy or non-state actors in order to achieve their goals. Frank Hoffman put together some possible terms to describe the 21st kind of war, such as Fourth Generation Warfare, Compound Wars, Unrestricted Warfare, and, of course Hybrid Wars, which he described as "blending the lethality of state conflict with the fanatical and protracted fervor of irregular warfare." (Hoffman, 2007).

At the same time, the European Parliament and the Council (2016) presents hybrid threats as a combination of classic and unconventional measures used in order to achieve some goal without declaring it a war, stating once again the combination of methods and forces. This fact is supported by Arsalan Bilal, too. He suggests that "hybrid warfare entails an interplay or fusion of conventional as well as unconventional instruments of power and tools of subversion. These instruments or tools are blended in a synchronized manner to exploit the vulnerabilities of an antagonist and achieve synergistic effects." (Bilal, 2021).

While the methods are considered to be old, what might count as new are the tools used. These new instruments, supported by the technological progress bear consequences on a deeper level than the conventional ones which impacted the physical and the material.

In addition, if we take a deeper look into the succession of events that took place, one notices that the conflict seems to be driven on multiple levels, military, political, economic or psychological. For the clarity of work, we will only focus on the psychological one which consists of propaganda, disinformation and other tools to deceive and manipulate through false information presented as true. This could also lead to fear, insecurity, polarization of the society, mistrust for the public institutions or the democratic process. In contrast to a classic war, the Russian-Ukrainian conflict did not begin with the first missile triggered, but with disinformation, propaganda and cyber-attacks. (The Spectator, 2022).

Physical harm is presented to us every day by mass-media tracking down the number of buildings, facilities, deaths and injuries. At the same time, little is said about the psychological damage, the traumas that people live every day. Regarding the Russian-Ukrainian war, one talk about the conflict's emotional consequences emphasizing the fear shared among people.

According to the Center for Strategic and International Studies, there are 7,5 million refugees who have fled Ukraine since February 24th and almost 7 million people who have been displaced around Ukraine (Yayboke, et al., 2022). Thus, migration is one of the fear-triggered actions taken by the Ukrainians in this ongoing conflict, but it could also be seen as a threat to neighboring states' national security. For instance, Romania is not a very welcoming country for refugees and immigrants, according to a INSCOP study concerning the refugees from Syria. Moreover, the data shows that the percentage of people who disagree with the immigrants coming into Romania increased with almost 20% during a few months, from September 2015 to march 2016, (INSCOP, 2016) in the context of the refugee crisis that EU was facing at the time and soon after the terrorist attack in Brussels.

As a specific feature of hybrid war, by targeting the vulnerabilities in any ways and with any means, the Russia seems to follow the rules in a tridimensional approach, targeting the critical infrastructure with missiles, the technologies with cyber-attacks and the civil population with disinformation and propaganda, all at the same time.

2. The Russian-Ukrainian hybrid conflict

The hybrid character of war is a not-so-new matter. Almost every component pre-existed, but the synchronized usage of new methods and means in a highly technological world is what gives the war even a more unpredictable nature and what challenges us the most. The novelty of hybrid war seems to be the same in Giannopoulos' (2021) and Galeotti's (2016) views as well.

One issue could be to determine whether the concept of hybrid war appeared to define the Russian type of war or it came to define the contemporary conflict characteristics and was later attributed to Russia's type of war. According to Fridman (2018), the concept was created to better define the contemporary conflicts but has reached to be like an exchange coin of intimidation and disinformation between Russian and the West on all levels of the society. In Russians approach, however, the term *gibridnaya voyna*, the translation for *hybrid war,* is used to define the threats and actions taken by the West against the Russia and doesn't explain the Russian strategy of war. Furthermore, we learn that the understanding of the concept hybrid war/*gibridnaya voyna* has different values in the West and Russian approaches, while they seem to be accusing each other of carrying out a hybrid war.

Even though the terms *hybrid war* and *gibridnaya voyna* seem different, we identified some key similarities between the two:

- The use of conventional and unconventional tactics combined
- The vulnerabilities' exploitation on multiple levels: societal, political, economic
- The blurring of the state boundaries
- The asymmetry of the forces

Starting with the Russian actions taken in Crimeea in 2014, which in the Western approach, qualify as hybrid war, the Russians intensified the use of hybrid methods against other states. For example, the actions taken in the Western Balkans, especially after the annexation of Crimeea fall into three areas, which, in Solik et al. view are: "political and economic influence, information space and proxy organizations" (Solik, et al., 2022), and qualify as hybrid threats. Another example, presented by Kragh and Asberg (2017) shows how Russia uses millitary threats, forgeries, disinformation and agents of influence in order to influence the public opinion in Sweden. The psychological warfare carried out by Russia ever since 2014 was analyzed by Veeble (2016), too, who focused his research on the Estonian case. Furthermore, we may not forget about all the cyber-attaks attributed to hacking groups which we believe to be founded by the Russian Government. A short presentation and comparison between three important attacks against Estonia in 2007, Georgia in 2008 and Ukraine in 2014, was made by Mihai (2022). Russia continued its cyber attacks in Ukraine for the whole period, but there was an increase in their in frequency and intensity as the day of the invasion approached. According to a Microsoft Report regarding the cyber component of the conflict in Ukraine, "The Russian invasion relies in part on a cyber-strategy that includes at least three distinct and sometimes coordinated efforts—destructive cyberattacks within Ukraine, network penetration and espionage outside Ukraine, and cyber influence operations targeting people around the world." (Smith, 2022) Russia is adapting the way it uses cyber-attacks depending on the immediate strategic objective. It uses destructive cyber-attacks in Ukraine to loosen their critical information infrastructure and cyber influence operations to spread disinformation and propaganda, which have been for decades part of their foreign policy. (Kragh & Asberg, 2017)

There are lots of examples when Russia has used different technics or platforms for disinformation or propaganda. For example, the use of repetitive words or phrases, a known feature of Russian disinformation, used by outlets funded by the government, have the potential of making a message more interesting and trustworthy, for instance - *Nazi* or *fascist.* There are almost 500 examples of the use of the words *Nazi* or *fascist* referring to Ukraine (EU vs DISINFO, 2022). Furthermore, Treyger et al. state that "the fullest display of Russian information warfare—with, almost certainly, the greatest impact—took place in Ukraine, where Russia integrated information and kinetic operations, using disinformation and propaganda both as strategic tools to shape political outcomes and as operational tools to undermine military morale" (Treyger, et al., 2022, p. 58), which illustrates that the coordinated effort of the Russia implicates along with military forces, information technology experts, but also state-funded outlets to achieve their goals. The ongoing war includes all we previously knew about the Russian type of information warfare.

The Russian-Ukrainian conflict seems to be neither conventional, nor new. As Mark Galeotti argued about Russia's specific characteristics in carrying on a war, "Russia's current style of war reflects reforms dating back to 2008 and policy discussions going back much further than that. What really has changed is the context in which old methods are being applied" (Galeotti, 2016), so, we could say that there is a pattern in Russia's operating style. While there are some similarities with the annexation of Crimea, the Russian invasion in Ukraine seems to have not gone as planned. "A variety of data indicate that the aggressor was counting on a blitzkrieg" (Kasianov, 2022), as it happened in Crimea, but that's not how it happened now. Quite the opposite, it got extended, inflicting more and more harm to Ukraine and Ukrainians.

While we support the controversial meaning and we don't agree with the novelty of hybrid wars, we find that the continuous existence of hybrid threats, in war and peace could generate long term consequences in different aspects of the society. Starting with the cognitive and emotional one (information warfare) and continuing with the economic or political aspects (cyber-attacks) which can very much determine the well-being of a state and its citizens.

3. The Russian-Ukrainian psychological dimension of the conflict

The psychological aspect of the war or the so-called *psychological warfare* is as old as the war itself. There is nothing new in the fact that commanders always try to strengthen the morale of their soldiers and demoralize the opponents. Although this is an ancient concept there are lots of new tools and innovative means used in order to obtain strategic advantages.

The common element between psychological and information warfare is that the use of information technology in this context is to disrupt, degrade, or destroy an adversary's ability to use information systems and networks, while with the psychological one, the psychological tactics are called upon to influence the behavior of an adversary or target population. If there seems to be no similarity between the two, they both involve the use of communication and information to influence individuals or groups and are frequently combined or used together in order to obtain strategic advantages. Furthermore, considering the tools used, sometimes information technology, the information warfare primary tool, is used in order to spread propaganda and disinformation. Such an example is largely explained by Pantserev (2020), who considers that the evolution of information technologies, especially those with an artificial intelligence component could lead to serious implications and even to a "global psychological warfare" due to its capacity of replicating human characteristics and using them in malicious actions.

Other voices may say that the outcome of carrying either an information warfare or a psychological one may be the same and the same can be said about the purpose. For example, Treyger et al. state that "Russian information warfare threatens to undermine the integrity of democratic processes, erode the belief in factual truths, and cause concrete harm with well-timed or sophisticated disinformation" (Treyger, et al., 2022), considering disinformation as a tool used in informational warfare. In addition, Guess and Lyons support this statement, too. They consider that "the terms *misinformation*, *disinformation*, and *propaganda* are sometimes used interchangeably, with shifting and overlapping definitions. All three concern false or misleading messages spread under the guise of informative content, whether in the form of elite communication, online messages, advertising, or published articles" (Guess & Lyons, 2020). There is also a distinction between the three notions in terms of purpose and intentions. Misinformation involves spreading false information without the intention of misleading or deceiving the recipient while disinformation is all about the intention of sharing false information in a matter that determines the reader or listener to believe. Last, but not least, propaganda involves

a mixture of true and false information with the primary focus on generating the desired belief or actions in accordance with the issuer's agenda.

There is a distinction to be made related to the psychological perspective, between psychological war and psychological operations, the two terms describe quite different types of actions, and to avoid confusion we need to mention the following: the first concept applies in times of war or crisis and focuses on propaganda or disinformation in order to influence or manipulate, while psychological operations (PSYOPS) are used as a support of military operations, at peace.

In the ongoing conflict between Russia and Ukraine, since there is a very long history of tensions, propaganda and disinformation are largely used. To begin with, President Putin presented himself and Russia as liberators when they decided to attack Ukraine with military and non-military tools, arguing that he had no other choice to defend Russia and its citizens. Moreover, in 2014 "Russia was obliged to protect the inhabitants of Crimea and Sevastopol", for "self-defense" (The Spectator, 2022). This is a classic example of how Russia justifies its actions using propaganda. This is not the first time when Putin used this kind of strong brotherhood allegations with the Ukrainian people to justify the offensive actions taken against Ukraine (Pynnöniemi & Rácz , 2016).

At the same time, some arguments brought by President Putin could have been accurate, as Jānis Bērziņš stated, too, "Since for Russia, Ukraine is supposed to be a close ally or, at best, neutral, […] the involvement of the United States and the European Union in Ukrainian internal affairs (are seen) to be a direct confrontation to its regional interests" (Bērziņš, 2014). This of course does not justify the hybrid actions lead by Russia against Ukraine, neither the military operations, nor the cyber-attacks or the disinformation and propaganda, but the military support from the West is confirmed and received by Ukraine since February, 2022. Even though there are no other states fighting on the battlefield against Russia, and the military support consists only in military equipment, money and training for both authorities and military personnel, President Zelenski asked foreigners to join them and there was created a website – Fight for Ukraine, for that matter.

During this period of time, from February till now, there were lots of press released from Russia's President using the state-founded and state-directed media such Russia Today or Sputnik as tools of spreading propaganda and disinformation. As a result of his actions for destabilization of Ukraine, the European Council decided to forbid the broadcasting or other related actions in any ways for a number of Russian funded outlets (European Council, 2022). Furthermore, according to the Global Engagement Center, there are five main pillars that Russian disinformation and propaganda system rely on:

- Official Government Communications;
- State-Funded Global Messaging;
- The Cultivation Of Proxy Sources;
- The Weaponization Of Social Media;
- Cyber-Enabled Disinformation (Global Engagement Center, 2022)

An example of how Russian official government communications could impact the lives of either Russians or Ukrainians, but also people from neighboring states could be Putin's announcement of force enrolling for young men in Russian Army on September, 21st. That generated fear among Russians as a large number of men left the country, leading especially through Georgia, Kazakhstan, Turkey, Israel and United Arab Emirates, while others seek for legal advice to avoid enlisting or are simply hiding (Faulconbridge & Liffey, 2022).

Another official government communication with great impact was Russian leader's speech on May, 9th at a military parade in Moscow's Red Square. Using disinformation and propaganda, President Putin tries to justify his actions, and also increase the level of support from both the Russians and other states. In addition, Fusiek stated that "by using humiliation narratives and evoking patriotic images, Putin exacerbates a sense of victimhood and historical right in the Russian audience that emphasizes the state's isolationism and historic role to fight against the "Nazified" West and Ukraine" (Fusiek, 2022) proving once again the amount of propaganda and disinformation in President Putin's discourse, which can lead to confusion, false perception on the facts and anxiety.

4. Emotional and long-term consequences

War is a devastating event that can have economic, political, social or psychological consequences for both the combatants and the non-combatants, but nowadays, the hybrid war seems to generate less deaths related to war and an increase in the number of people who suffer from mental health disorder, as a consequence of war

(Rozanov, et al., 2018). The Russian-Ukrainian conflict makes no exception in terms of psychological damage caused to the Ukrainians, but also to people from neighboring states.

The psychological consequences of war have been studied over the past years with foreseeable results. It looks like, even after many years, the post-traumatic stress disorder or anxiety are still affecting the respondents as showed by Priebe et al. (2012) in their study about long-term mental health consequences of war conducted eight years after the war in former Yugoslavia. At the same time, during a conflict, people may have different kind of emotional reactions, such as flashbacks, nightmares and avoidance behaviors. Additionally, individuals may experience depression, anxiety and other emotional disturbances. More recent studies, conducted on the changes in mental health among Ukrainian due to the Russian-Ukrainian conflict show that there was an increase anxiety, depression, loneliness and hazardous drinking (Hyland, et al., 2022), (Shevlin, et al., 2022).

The first emotion triggered when an event of such magnitude as a war happens is fear. From a psychological point of view, the behavioral responses to fear are what they call the Fight-Flight-Freeze System (Donahue, 2020), where fight is the aggressive body's response to a threat, flight represents the urge to leave from the threat and freeze, the impossibility to move or to react in any way. Of course, this are the body's primarily reactions but we believe that it could be extended to Ukrainians' actions due to the Russian invasion of Ukraine as we witnessed some enlisting in the Army and fighting against the Russians, some leaving the country through mostly European countries and some others doing nothing, hiding in the shelters provided by the state and praying for their lives.

The Russian-Ukrainian war outbreak, coming right after almost 2 years of Covid-19 pandemic, had an overwhelming impact on the health care system, which, besides the actual work of providing care, had lots of their facilities directly attacked and, on the Ukrainians' mental health, which had already been affected by the ongoing pandemic (Chaaya, et al., 2022). The Ukrainians basic needs were threated, such as safety, shelter, electricity or communication. The communication disturbances that occurred due to a cyber-attack against Viasat's KA-SAT network were attributed to the Russia's funded hacking groups (Viasat, 2022). In this context, one of the Ukrainians' reactions was leaving the war area. Some of them, more than 7 million left the country, while almost the same number of people were internally displaced. (Yayboke, et al., 2022)

At the same time, while the consequences for those directly affected by the war were long studied, little is said about the people from neighboring states or even living further apart from the war. The immediate access to information allows us to be in touch with any event that happens anywhere on the globe. The atrocities that continue to happen in Ukraine, indispensable in any news report and from any social media platform impact the lives of millions. The fact that something like that could happen in our era is perceived by some as totally unpredictable, unjustified, horrific and frightful. Moreover, even Russians faces the consequences of the sanctions applied to their own country. Among the consequences one mentions: the reduction on medical supplies, the financial difficulties, the interdiction to participate in certain events, to which the almost unanimous foreign opinion that any Russian civilian can be held responsible for Russia's actions is added. It is likely that all of these cause serious mental health challenges to the Russian population (Volunteer FDIP, 2023).

The psychological trauma was the first consequence felt, people started fearing for their lives and their loved ones, war proximity being an important factor that they considered, especially if we talk about the Republic of Moldova's population, Romania or Poland. For example, people in Romania considered to update their travel documents, to exchange their money into euro or to put gas into their cars in order to be prepared to flee if the war area extended (Hațegan, 2022).

The next consequence that emerged was the wave of refugees coming from Ukraine. According to the UN Refugee Agency (2023), there are 1.5 million refugees who have registered for support in Poland, over 100000 in Romania, almost 500000 in Czech Republic, and there are a few other countries where refugees arrived and seek for help. These waves of refugees laid tremendous financial and social pressure on the countries receiving them, in incipient stages. However, studies show that the refugees could be an opportunity for the host country's economy (Taylor, et al., 2016) as a long term desiderate. On one hand, the state, along with the large number of volunteers, seem to have contributed positively by helping Ukrainian refugees with social assistance, shelter, food, water, medical supplies and health care (The Romanian Government, 2022), but, on the other, being so involved and emotionally close to the Ukrainian refugees, it is possible that the volunteers have faced difficulties with their mental health, too, such as burnout and secondary traumatic stress, as they were exposed to the traumatic experiences of the refugees (Roberts, et al., 2021).

5. Conclusions

Wars have always been destructive and the damage created, either material or psychological, has been difficult or even impossible to manage or repair. The hybrid wars make no exception as they are a combination of classic military actions and non-linear operations, the asymmetric character creating even more difficulties in identifying all the actors involved or the tools used. There is more uncertainty and unpredictability than in a classic war which generate greater issues in dealing with the emotional and psychological consequences.

The Russian-Ukrainian conflict seems to qualify as a hybrid war as there were all kind of hybrid threats materializing during the last year in the war area and beyond. Before the outbreak of the war and continuing after, Russia used tools like disinformation, propaganda, cyber-attacks in order to influence and manipulate the public opinion into giving them their full support and making them believe that its actions are justifiable. And for a short period of time it worked, at least for the Russian population as a study conducted on the Russian populations shows that over 80% support the decision of attacking Ukraine (Yuri Levada Analytical Center, 2022).

The use of psychological methods and tools, such as disinformation or propaganda have the potential of creating a state of panic, insecurity, and various types of mental health disorders, such as anxiety or depression. All these consequences we found out to emerge in the Russian-Ukrainian hybrid conflict.

This paper identified the possible emotional consequences of the ongoing war in Ukraine, emphasizing the mental health issues identified in the Ukrainian population due to exposure to traumatic events, but also in neighboring states' population which was not, in any way, sheltered from the events happening at their states' borders, especially due to the rapid spread of information in the online environment.

References

Bērziņš, J., 2014. Russia's New Generation Warfare in Ukraine: Implications for Latvian Defense Policy. April.

Bilal, A., 2021. Hybrid Warfare – New Threats, Complexity, and 'Trust' as the Antidote. *NATO Review,* 30 November.

Chaaya, C. et al., 2022. Ukraine – Russia crisis and its impacts on the mental health of Ukrainian young people during the COVID-19 pandemic. *Annals of Medicine & Surgery,* July.

Donahue, J. J., 2020. Fight-Flight-Freeze System. In: *Encyclopedia of Personality and Individual Differences.* s.l.:s.n., pp. 1590-1595.

EU vs DISINFO, 2022. *Key Narratives in Pro-Kremlin Disinformation: "Nazis".* [Online]
Available at: https://euvsdisinfo.eu/key-narratives-in-pro-kremlin-disinformation-nazis/#.

European Commission, 2016. *Joint Framework on countering hybrid threats - a European Union response,* Brussels.

Faulconbridge , G. & Liffey, K., 2022. *Reuters.* [Online]
Available at: https://www.reuters.com/world/europe/russia-says-over-200000-drafted-into-army-since-putins-decree-2022-10-04/.

Fridman, O., 2018. *Russian "Hybrid Warfare". Resurgence and Politicization.* s.l.:Oxford University Press.

Fusiek, D. A., 2022. Putin's Great Patriotic War: Russia's Securitization of the West and Humiliation Narratives Surrounding the 2022 Invasion of Ukraine. *HAPSc Policy Briefs Series,* pp. 105-113.

Galeotti, M., 2016. Hybrid, ambiguous, and non-linear? How new is Russia's new way of war?. *Small Wars & Insurgencies,* 22 03, 27(2), pp. 282-301.

Giannopoulos, G., Smith, H. & Theocharidou, M., 2021. *The lanscape of hybrid threats: A conceptual model.* Luxembourg: Publications Office of the European Union.

Global Engagement Center, 2022. *Kremlin-Funded Media: RT and Sputnik's Role in Russia's Disinformation and Propaganda Ecosystem*.

Guess, A. M. & Lyons, B. A., 2020. Misinformation, Disinformation, and Online Propaganda. In: *Social Media and Democracy*, pp. 10-33.

Hațegan, O., 2022. *How Romanians are preparing to flee the country in case the war in Ukraine spreads. Queues at the passport offices, savings exchanged in euros, powers of attorney for travelling with minors and leased cars*.

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

Hyland, P. et al., 2022. The psychological consequences of war in Ukraine: Assessing changes in mental health among Ukrainian parents. *PsyArXiv,* 07 October.

INSCOP, 2016. *April 2016 – Refugee crisis.* [Online]
Available at: https://www.inscop.ro/aprilie-2016-criza-refugiatilor/
[Accessed 28 november 2022].

Kasianov, G., 2022. From Historical Fallacy to Tragic, Criminal Loss: Putin's Case for Invading Ukraine. *Kennan Cable,* Aprilie.Volume 76.

Kragh, M. & Asberg, S., 2017. Russia's Strategy for Influencing throgh Public Diplomacy and Active Measures: The Swedish case. *Journal of Strategic Studies,* 05 January, 40(6), pp. 773-816.

Mihai, P., 2022. Provocări hibride de natură cibernetică. *Infosfera,* Issue 2.

Pantserev, K. A., 2020. The Malicious Use of AI-Based Deepfake Technology as the New Threat toPsychological Security and Political Stability. In: *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity.* s.l.:Springer Nature Switzerland AG, pp. 37-53.

Priebe, S. et al., 2012. Psychological Symptoms as Long-Term Consequences of War Experiences. *Psychopathology,* Issue 46, pp. 45-54.

Pynnöniemi , K. & Rácz , A., 2016. *Fog of Falsehood. Russian Strategy of Deception and the Conflict in Ukraine.* Helsinki: The Finnish Institute of International Affairs.

Roberts, F., Teague, B., Lee, J. & Rushworth, I., 2021. The Prevalence of Burnout and Secondary Traumatic Stress in Professionals and Volunteers Working With Forcibly Displaced People: A Systematic Review and Two Meta-Analyses. *Journal of Traumatic Stress,* 27 March, 34(4), pp. 773-785.

Rozanov, V. et al., 2018. Mental Health Consequences of War Conflicts. In: *Advances in Psychiatry.* s.l.:Springer, Cham, pp. 281-304.

Shevlin, M., Hyland, P. & Karatzias, T., 2022. The psychological consequences of the Ukraine war: What we know, and what we have to learn. *Acta Psychiatrica Scandinavica,* pp. 105-106.

Smith, B., 2022. *Defending Ukraine: Early Lessons from the Cyber War*.

Solik, M., Graf, J. & Baar, V., 2022. Hybrid Threats in the Western Balkans: A Case Study of Bosnia and Herzegovina. *Romanian Journal of European Affairs,* 22(1).

Taylor, E. J. et al., 2016. Economic impact of refugees. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES,* 05 July.113(27).

The council of the European Union, 2022. *COUNCIL DECISION (CFSP) 2022/351 of 1 March 2022 amending Decision 2014/512/CFSP concerning restrictive measures in view of Russia's actions destabilising the situation in Ukraine,*.

The council of the European Union, 2022. *Răspunsul UE la invadarea Ucrainei de către Rusia.* [Online] Available at: https://www.consilium.europa.eu/ro/policies/eu-response-ukraine-invasion/.

The Romanian Government, 2022. *Facitities for Ukrainian Refugees.* [Online] Available at: https://www.gov.ro/ro/pagina/facilitati-pentru-refugiatii-din-ucraina.

The Spectator, 2022. *Full text: Putin's declaration of war on Ukraine.* [Online] Available at: https://www.spectator.co.uk/article/full-text-putin-s-declaration-of-war-on-ukraine/

The UN Refugee Agency, 2023. *Ukraine Situatian Flash Update #38*.

Treyger, E., Cheravitch, J. & Cohen, R. S., 2022. *Russian disinformation efforts on social media.* Santa Monica: RAND Corporation.

Veebel, V., 2016. Estonia confronts propaganda: Russia manipulates media in pursuit of psychological warfare. *Per Concordiam,* 7(1), pp. 14-19.

Viasat, 2022. *KA-SAT Network cyber attack overview.*

Volunteer FDIP, 2023. *Ukraine-Russia War and Its Impact on Psychology and Mental Health Globally.* [Online] Available at: https://www.volunteerfdip.org/ukraine-russia-war-and-its-impact-on-psychology-and-mental-health-globally.

Yayboke, E., Strouboulis, A. & Edwards, A., 2022. *Update on Forced Displacement aroud Ukraine* [Interview] (3 October 2022).

Yuri Levada Analytical Center, 2022. Approval of Institutions, Ratings of parties and Politicians.

Target Audiences' Characteristics and Prospective in Countering Information Warfare

Daniel Ionel Andrei Nistor
National Defence University „Carol I", Bucharest, Romania

dan.nistor.rp@gmail.com

Abstract: NATO Defense Education Enhancement Program defines Information Warfare as an operation run to get cognitive assets over the opponents, by controlling one's own information space while disrupting the opponents' one. Not new as a process, continuous technological progress has endowed this phenomenon with speed and instruments to fight cyber and cognitive battles, to attack perceptions, trust, polarise and disrupt societies at large. The all present and undergoing kinetic conflict between Russia and Ukraine doubled by an even stronger cognitive and information war since February 2022 has highlighted even more the need to better understand individuals' behaviour and characteristics when faced with unconventional attacks, irrespective of a passive or active feedback. By identifying and analysing specific public categories, one can establish which are contextual variables that trigger a social reaction, to be able to then design a set of protective or defensive measures. For a full understanding of the way Information Warfare impacts people's thinking and decision-making process, to see how a resilience plan can be designed, one should investigate not only the information war instruments but also their effects over people at large. Not knowing the voice of the hostile authors, it Is still important to understand the domestic audience and their reaction to it, so that protective actions be taken for resilience and protection, through education. The domestic public's identity and its dominant characteristics are brought into attention to understand which is the relation between these and the way Information Warfare can be countered through education, with examples from the Russian's hostile activity. Values, national identity, stereotypes and generalist psychological profiles will be looked at in this paper, to be put in relation to behaviours, attitude change and resistance in front of types of messages, campaigns and types of media-embedded grey zone threats. The present paper is part of a larger PhD research program that focuses on consolidating a society's security culture through better institutional strategic communication, therefore all the findings will be used to this end.

Key words: Information Warfare, Disruptive Process, Behaviour Change, Target Audience

1. Introduction

In order to counter Information Warfare, is It key to anticipate and plan accordingly - because many scenarios are mostly recurrent and repetitive, or is it necessary to change the approach in order to better understand the situation and act to prevent it long before it occurs? It sounds like being a choice of creating awareness and not dealing with consequences. Can growing awareness change or alter the degree in which we are affected by Information Warfare? We will try to provide an answer to these open questions.

In this research paper we will investigate how values, national identity, stereotypes and individuals' psychological profiles can be put in relation to behaviours, attitude change, resistance to propaganda and influencing actions to create structural changes within a specific society.

Yet, the purpose of this article is not to analyse the evolution or the techniques, tactics and procedures of Information Warfare, but rather to forward an approach that focuses on the audience, as countering Information Warfare is not only about doctrine, strategy, techniques, tactics and procedures, but also about audiences and a larger societal context.

Therefore, we need to examine how our audience looks like, what we want to protect, how we can understand the context, and how to use the proper channels to change behaviours, including but not limited to educating the audience and teaching them how to protect themselves and others or be part of the process.

Not more than half a century ago, Information Warfare was an attribute associated predominantly with the military and, to a lesser extent, with the political. The war of public statements was the main instrument that kept political figures on top of audiences' mind. To the same extent, public statements could make an official vanish from the political stage overnight, either following a scandal or by a simple twist of words.

Today, Information Warfare is present everywhere - from social to economic, military or political, and its impact is often difficult to forecast or anticipate. Big data and the massive flow of information have fundamentally changed the way media reports. More than half of the world's population has access to Internet and, in a single day, more than 333 billion emails, 650 million tweets and 41,6 million WhatsApp messages are sent; 150K messages and 147K images are sent and received on Facebook, while Instagram users post around 347K stories

per day. (Bulao 2023) Taking into account that almost 60 percent of the total global population has a social media account, one can easily understand the dimension of this phenomenon (Datareportal, 2022).

Whether we are talking about sovereign states, state actors, alliances, coalitions or private entities, we must be aware of the fact that the techniques, tactics and procedures have been largely adapted to the environment and context. As a result, any protective strategy (process) should be rethought and refocused from channels to the audience we want to protect. At the same time, we ought to reanalyse if the doctrine should be redone or just readapted, in order to see if protecting the public could be seen as a strictly defensive operation or it can also involve offensive actions. What are the limits of Information Warfare countermeasures? The areas of Information Warfare, its methods and channels have multiplied and diversified, making it more and more important to understand all these developments in the context of understanding the audience.

2. Information Warfare and the Security Ecosystem

According to Sartonen (2022) the concept of "war" is being transformed due to in the inclusion of Information Warfare in the digital environment. The cyber domain and the internet in particular allow global audiences to be reached in battles of influence (all, 2017), pressuring us to seek to understand the "war" in a broader ecosystem. As Severinghaus (2019) underlines, we are operating with seven warfare areas: the first five are the domains traditionally associated with the military (air, sea, land, space, and cyber), while the other two, not explicitly recognised as such, are economic and education warfare. "If one wants to be politically correct, economic and education warfare are characterised simply as "competition". But, in truth, other nations of the world, not just China, are stealthily waging economic and education warfare outside the military, and others are significantly outpacing the United States in educational achievements in many areas with national security implications". (Severinghaus 2019)

NATO Defence Education Enhancement Program defines Information Warfare as an operation ran to get cognitive assets over the opponents by controlling one's own information space and disrupting the opponents'. In a conceptual delineation, Dan Kuehl defined Information Warfare as "the conflict or struggle between two or more groups in the information environment", while professor Stupples sees Information Warfare as "the integration of electronic warfare, cyberwarfare and psychological operations, for both attack and defence". (Stupples, 2015). To understand Information Warfare, we should therefore explain the larger framework of warfare and analyse why it is different today, compared to two decades or a century ago. Western leaders are investing billions to develop capabilities matching those of China and Russia, establishing military commands for attacking, defending and exploiting the vulnerabilities of electronic communications networks. Information Warfare combines electronic warfare, cyberwarfare and psychological operations into a single fighting organisation, and this will be central to all warfare in the future (Stupples, 2015).

The concepts of "security" and "securitization" were largely studied and explained by Buzan (1991) who defines security as "the pursuit of freedom from threat and the ability of states and societies to maintain their independent identities and their functional integrity against forces of change which they see as hostile. The bottom line of security is survival, but it also reasonably includes a substantial range of concerns about the conditions of existence. Quite where this range of concerns ceases to merit the urgency of the "security" label (which identifies threats as significant enough to warrant emergency action and exceptional measures including the use of force) and becomes part of everyday uncertainties of life is one of the difficulties of the concept". (Buzan, 1991)

Influencing audiences and changing behaviours are just a part of Information Warfare. The entity, state or state actor are executing Information Warfare actions on a targeted audience to change perceptions or behaviours in order to achieve a specific objective. The main goal is not to change the behaviour, but to reach an objective. This is the main reason why audiences should be the point of focus when conducting Information Warfare or in countering hostile Information Warfare. Even if we are talking about changing behaviours in order to accept a new political regime, a new social policy, new state borders or significant economic changes, the hearts and minds of the audience are more efficient and enduring than using physical force. Using "supreme excellence in breaking the enemy's resistance without fighting" (Tzu 2011) has always been the most powerful way of achieving an objective.

The struggle to move confrontation from physical to informational, political or economic areas while reducing loss of life as always been constant aim of states, alliances and organisations. Hence, Information Warfare can be construed in relation to two possible approaches: "as a movement from the area of hard power to that of soft power, and as a mutation of the confrontation having the centre of gravity not in the physical domain but

in information domain (using the cyberspace and electromagnetic spectrum)". (Lesenciuc 2016) Information Warfare can be used by states in foreign policy to employ and advantage. (Lewis, 2022)

Besides, Information Warfare is not a new phenomenon, yet it contains innovative elements such as technological development effects, resulting in information being disseminated faster and on a larger scale. Ullman highlights a series of factors that can change the world order by simply influencing audiences: "*failed or unperformed governance, climate change, cyberspace, social networks, drones, terrorism and state debt.* Seen as a new doctrine that marks the transition *from a Cold War doctrine based on nuclear weapons only, we are now witnessing the creation of a new doctrine that has the same mechanism of action as a virus: no physical border or internal or external state policy, known as transition from MAD (Mutual Assured Distraction) to newMAD (Massive Attack Disruption)".* These are all disturbing and disruptive factors for the world order, acting independently or simultaneously, charged with huge potential for disruption or destruction (Ullman, 2021). Each of those factors can become either a channel or an influencing factor that, attuned to the right massage, can reach and influence the targeted audience in order to achieve the Information Warfare objectives.

Information Warfare and propaganda have always been outstanding instruments for Russia. The Kremlin used its powerful narrative in most of the Baltic, former Communist or Soviet states, including Czech Republic, Slovakia, Poland, Hungary, Ukraine, Bulgaria, Georgia, Moldova and Romania. Before the Russian invasion in Ukraine, the authorities were scarce in creating coherent strategic plans or in joining efforts to counter Kremlin's influence, being mostly concerned with domestic political issues or worse, being completely unaware of the imminent risk. Russian Information Warfare in Black Sea Region was present long before the attack on Ukraine, in February 2022. The Kremlin has been playing strong since the annexation of Crimea in 2014. (Lucas, 2014). The "green man's" presence in Eastern Ukraine was just the tip of the spear; Moscow targeted the Russian-speaking minority and conducted military exercises in the Black and Baltic Seas to project the image of a military superpower. Thousands of cyberattacks have been executed, severely affecting critical infrastructure. The Ukrainian electricity grid was targeted in 2015, leaving more than two thousand people with no power, followed by an attack on Kyiv in 2016 that generated a major blackout. (Scutaru, 2022)

To counter Russian Information Warfare, NATO, the EU and Ukraine developed programs in all domains - from political to military, economic and social. They imposed diplomatic and economic sanctions, closed and banned Russian media outlets, protected their own cyberspace and physical borders. The rules of conducting propaganda activities on social networks have been also updated, becoming more restrictive, educating and promoting awareness. NATO readjusted its Strategic Concept, strengthened his Eastern and Northern flanks, deployed the Battle Groups at NATO's borders and in the countries neighbouring Russia, and increased military cooperation with Ukraine, Georgia and Moldova by conducting military excises and training missions. Western governments pressed social media firms to remove pro-Russia propaganda after the war in Ukraine began, and outlets like RT (Russia Today TV station) have been wiped from popular platforms. That may create an impression in much of the world that related disinformation is on the decline. (Letzing, 2022)

3. The Role of Audience in Information Warfare

In order to protect the audience and counter Information Warfare, we need to understand the actors, the environment and how the "game" is be played. The effectiveness of our actions depends on understanding audiences, the context, instruments and adversary objectives, while all of those measures are executed from the defensive strategy, aiming to prevent malign influence and protect domestic audiences. At the same time, another way of countering Information Warfare is by offensive actions, targeting the adversary in order to diminish, reduce or even annihilate his capability of action.

In Informational Warfare, the logical scheme is defined by the relation between two actors, where actor A has a clear mission and Actor B has the objective of not letting Actor A fulfil its mission. In this relationship, the public is not a target *per se.* Yet, protecting its people is the most powerful way through which Actor A can fight against the mission of actor B, not allowing him to accomplish its objective. Developing further on this scheme, Actor A targets Actros's B audience to change its behaviour in order to accomplish its objectives. Actor B, at the same time, needs to protect its own audience from Actor A's malign influence. In protecting its audience, Actor B can execute defensive actions as pre-emptive actions, protective actions or offensive operations. In today's world, we face a multitude of potential influencing factors that both actors A and B can make use of to influence audiences in Information Warfare. A psychological operation has largely three major steps: (1) establish the objectives and define target audiences (TA); (2) produce, approve and deliver to the TA the influence messages; (3) assess the effectiveness; (FM3-05.302, 2005). To maximize the results, it is important to disseminate the

influencing messages and proper evaluate their effectiveness to the same extent as it is key to focus the message on behavioural change and execute all the required actions.

Individuals tend to focus more on values, certain people, objects, phenomena, and situations that can usually satisfy their needs, being highly influenced by previous experience, tradition and cultural characteristics. Yet, in order to determine an individual's behavioural patterns, we must therefore consider first how the *interest, values, social norms and social sanctions* are adopted through the process of socialisation (AJP-3.10.1(A) Allied Joint Doctrine for Psychological Operation 2009). Groups act different than individuals, and are largely characterized by "impulsivity, mobility and irascibilities of the masses, animated and manipulated by leaders who understand the power of suggestion and the gullibility of the masses." (LeBon, 2022)

4. Instruments and Research Methods

When analysing the "Romanian case", we used observation and comparative analyses as research methods. For comparative insight, we turned to Hofstede's 6D cross-cultural model, showing the effects of a society's culture on the values of its members, and how these values specifically relate to behaviour. Hofstede's study covers 70 countries and has been last updated last in 2010 (Hofstede, 2006)**.** In addition to Hofstede's study, we used David's research (2015) "Psihologia poporului roman" (The Psychology of the Romanian People), a comprehensive cultural study forwarding the psychological profile of the Romanian people from a cognitive-experimental perspective. One of the central points of David's study is the comparative analysis of how Romanian's see themselves, how they are seen by others, and how they actually are. (David, 2015)

For clarity and scientific relevance, we have followed up Hofstede's and David's studies, which provide a revealing information about the Romanian society from a psycho-social point of view, with two security barometers, realized by the Strategic Thinking Group, entitled "Public distrust: West vs. East, the Rise of Nationalism in the Disinformation Era, and the Fake News Phenomenon", (StrategicThinking, 2022) and respectively with a barometer called "The Security of the Future" realized by the Romanian National Institute of Statistics". These two barometers shed some light on the feelings and perceptions of Romanians regarding their trust in the military and political European and Euro Atlantic alliances.

And last but not least, for the particular investigation on how Romanians responded to fake news and disinformation immediately after the start of the conflict in Ukraine we referred to a study called "Propaganda Without Borders, A study of pro-Kremlin propaganda among far-right and radical voices in Hungary, Poland, Romania and Serbia" carried out by Global Focus, an NGO specialised in analysing pro-Russian propaganda in Romania and Eastern Europe. (GlobalFocus, 2022)

To end with, the observational dimension of this paper is supported by a number of narratives retrieved from various Romanian media outlets in 2022, in the context of the Ukrainian conflict, aiming at emphasising the relation between the characteristics of the Romanian audience and their reaction to hostile narratives.

5. The Romanian Case

To illustrate how values, national identity, stereotypes and psychological profiles can be put in relation with behaviours, attitude change and resistance to messages, campaigns and various types of media-embedded grey zone threads, in the larger contexts of Information Warfare, we shall conduct an analysis on Romanian audiences.

The first four narratives that took shape in the immediate aftermath of Russia's invasion of Ukraine in Romania are: (1) related to the food crisis, a scenario in which s*hops are running out of basic goods, such as sunflower oil, flour and sugar, because Ukraine was one of the biggest producers in the world; food will increasingly become more and more expensive; therefore, we need to make supplies;* (Antena3, 2022) (2) The rising price of petrol - *the price will dramatically increase (with 50% or more); gas stations become crowded overnight with drivers willing to purchase gallons of fuel before prices spiked.* (EuropaLibera, 2022) (3) Chaotic purchase and unsound administration of iodine pills under the immediate threat of Russia employing nuclear assets (Munteanu, 2022). And last but not least (4), *inflation*, *cash shortage or lack of money in ATMs; for fear of not losing money due to raising inflation, people need to withdraw cash immediately*. (RomaniaLibera, 2022). The response of Romanian citizens was purely emotional, creating chaos and extra demand that temporarily resulted in increasing the price of various goods. People created huge queues at gas stations and ATMs to make supplies and safeguard their savings. The state institutions promptly responded and managed the situation reassuring Romanian people and conveying clear messages to counter fake news.

The *distribution and acceptance of power and authority* and a high preference for avoiding uncertainty (Hofstede, 2006) are the psychological traits that pushed Romanian people to respond to messages that highlighted a potential, urgent threat, also explaining the chaotic, emotional response. Romanians are still fighting a cultural trauma related to lack of food, fuel and closed borders, dating back to the Communist era. At the same time, Romania is considered to be a collectivistic society (Hofstede, 2006), characterised by "*long-term commitment* to the member groups", which could be the family or extended relationships. Loyalty is also highly important, as well as the relation between employee and employers. The West (meaning the European Union, United States or NATO) is where Romania should be heading politically, economically and military is what 77% of the respondents answered, while only 10.4% believe that Romania should head towards the East (nr. Russia, China). An important 12.6 percent of the interviewees chose not to respond**.** (StrategicThinking, 2022). *The long-term orientation* (Hofstede, 2006) maintains an equilibrium between tradition and the modern way of preparing for the future; the resistance to change is somehow moderate, closing the gap between the radical changes that took place after the fall of the Communist regime.

Romanian people are also extroverts, scoring high on amiability. The psychological threats related to family and community care (David, 2015) generated solidarity with the Ukrainian people in terms of providing safe passage, shelters and food for more than 3.3 million refugees (as retrieved on 12th of January 2023) and assisting with transportation of cereals and materials from Ukraine to Europe (Government, 2023). This relates to Information Warfare by proving that the Romanian people's embedded values are stronger than nationalist propaganda messages, backing the "toxic pacifist" idea that Romanian should step aside and support peace by not getting involved in any way (GlobalFocus, 2023). At the beginning of the war, Romanians strongly backed Ukraine, aside from the supply of military aid. According to a March 2022 national poll, 81% of respondents supported receiving refugees and 79% supported sending humanitarian aid (GlobalFocus, 2023). Romanian citizens' continuous efforts to support Ukrainian refuges are consistent with David and Hofstede findings about family care. Romanians provided not only food and shelter, but also access to education Ukrainian and the possibility for Ukrainian adult population to be employed. These attitudes are once again supported by David's study, according to which "minorities are better integrated in Romania, in comparison with Serbia, Slovakia or Ukraine (David, 2015). Censorship, propaganda, diminishing or restricting freedom of opinion and speech narrative (Global Focus Report 2022) match Romanians' uncertainty avoidance, individualism and power distance (Hofstede, 2006). Like a double-edged dagger, the public asks Government officials to take cautionary measures and counter propaganda, while also accusing public institutions of manipulation and censorship. L*eadership is transactional* (David, 2015) and the "people" will always feel the need to criticise, request urgent solutions and try to preserve the sentiments of freedom and democracy. The freedom of speech is one of the most important rights for Romanian people, as important as family and propriety. Reinstating the mandatory military service in the context of war (G4Media, 2022) was also a narrative easy to debunk. The conscription service has been suspended in peacetime since 2007, and as per David's findings, *Romanian's have a strong sentiment for fighting for defending the country, thus being in a strong opposition with 31 modern cultures* (David, 2015). The official position of the Minister of Defence was prompt and properly pushed on traditional channels, new media and social media. (i.e. there is a designated Ministry of National Defence website for fake news debunking called inforadar.) (Inforadar, 2023). Regarding their exposure to fake news and disinformation 54,7% of Romanian appreciated that they were highly exposed, while 41,8% consider their exposure low exposure or none. (StrategicThinking, 2022)

Another idea backed by nationalist groups is that of Romania being drawn into the conflict, risking a direct attack from Russia. Yet, as the low score on individualism and trusting the Western alliances diminished the impact of this right-wing propaganda, that failed to create a pro-East, anti-West cleavage within the Romanian society. Good governmental communication and protective measures helped keep this narrative at bay. The same explanation also applies to the "it's the refugees' fault" narrative, also pushed by nationalist propaganda to support the idea of rising prices for basic goods and medication, and failing of Romanian government to support the disadvantaged.

To sum up, understanding the friendly, tolerant, emotional, religious nature of of Romanian people, their positive, but sceptical personalities and rather undisciplined behaviour (David, 2015) proves to be highly relevant in understanding what type of messaging Romanian audiences can successfully be targeted with, in the larger context of Information Warfare, also offering strong insights for how to effectively respond, counter and educate.

6. The Social Media Impact

Media plays a crucial part in Information Warfare, providing the channels for both influencing through hostile propaganda or malign information and counter influencing through informing, educating and debunking. **Information Warfare is the battle to** "persuade and induce the sympathy of potential allies; and simultaneously spread confusion, uncertainty and distrust in the enemy's population". (Pelletier, 2022)

Social media expansion facilitates interconnectedness and access to information, thus amplifying social risks and vulnerabilities. Social media creates bubbles, and "what's in your filter bubble depends on who you are, and it depends on what you do. But the thing is that you don't decide what gets in. And more importantly, you don't actually see what gets edited out" (Gould 2019). By using social media, we can easily convert an audience. Sending messages is cheap and the distance between the message and the audience is reachable within a click. On social networks we can create groups and communities, support social causes and sell products or services, and more importantly, we can polarise, segment and split different categories, according to our interest. For example, in two weeks, the leader of the Romanian extremist party, AUR (The Alliance of the Union of Romanians), George Simion, generated over half million intersections (almost double than mainstream online media overall) on the topic of Romania's Schengen acceptance. (GlobalFocus, 2023)

Meta platform counts "5.3 billion global Internet users", while 3.71 billion people use at least one Meta app - Facebook, Instagram, Messenger, or WhatsApp every month. This means that "70% of all Internet users are active on at least one Meta app", but the truth is that many use more than one. Yet, the impact in not only measured in the large number of users, but also in the massive spread of data, with over "1 billion stories posted every day across Facebook and a potential advertising reach of 2.08 billion people" (Newberry, 2023). Now, imagine the targeting process in terms of speed, respectively how fast you can reach a larger audience using short message. If a malign actor passed on a message in large private groups and the message is then redistributed in even larger public groups and pages, it becomes almost impossible to identify the original source, let alone to map the spreading. As a result, any attempt the correct the messaging and address the disinformation becomes, to say the least, chimerical. If you are not in the "right" bubble, you don't stand a chance to locate the source of the deception.

"Social media is used to create accounts and develop groups that, **collaborated** with the algorithms behind social networks and other programming mechanisms developed using artificial intelligence such as bot networks or the construction of non-existent human profiles, ensure their transmission and support of messages, which, if necessary, can be used in the command and control of Information Warfare." (Preda, 2021) "*The Facebook data breach wasn't a hack. It was a wake-up call"* is one conclusion related to the Cambridge Analytica scandal, one of the first massive scale use of users' data without consent, aiming at manipulating the socio-political system, influencing decision and changing social behaviour. 87 million Facebook user had their data exposed and used by a political consulting firm engaged in an electoral campaign, using an application build in tool. (Romano 2018)

Social media has become a powerful instrument that can be easily used by any actor who aims at creating beliefs and changing behaviours. Is can be used offensively in Information Warfare to directly address the audiences with malign intentions, and defensively with purpose of educating and protecting.

7. Conclusions

In this paper, we analysed the relation between the psychological profiles, behaviours, attitudes, values, national identities and stereotypes of Romanian audiences, and the main disinformation narratives employed by malign actors, domestic and foreign, with the goal of influencing and disrupting the Romanian socio-politic environment, in the larger context the Russian-Ukrainian conflict. How Romanians reacted to these disruptive narratives correlates with their psychological and cultural traits, proving why knowing the audience remains a key element in Information Warfare. Success or failure is dictated by the degree in which the coercive message matches the cultural values, psychological profile and social norms of the targeted audiences. To the same extent, the degree in which social and state actors succeed in tailoring their protective measures and further educating their national audiences determines who gains the competitive advantage in influencing and changing behaviour.

The disinformation narratives could be also found in neighbouring countries, such as Bulgaria, Czech Republic, Poland, Serbia and Hungary, and even if the public response was somewhat similar to that to Romanians', the differences stem from the deep embedded cultural straits. For instance, if we examine the growing extremist-nationalist current in all the above-mentioned countries, we can observe that the affiliation is lower in Romania,

mainly due to the strong collectivism factor (70), uncertainty avoidance (90) and power distance (90), compared to Hungary, a country scoring lower on the collectivism (20), uncertainty avoidance (82) and power distance (46).

Of course, using this analysis we cannot predict behaviour, but we can anticipate action by mapping the activities, attitude changes and eventually put them into context. We can, however, set up alternative courses of action and reduce the number of scenarios based on behaviour changes. At the same time, we must own the fact that the malign actors have the same knowledge and understanding, and we must learn how to think and act "out of the box" to be one step ahead protecting our audience and countering the malign influence.

Social media, used as Information Warfare channel, remain one of the most cost efficient and effective tools for influencing and protecting the audiences. There are more than 5.3 billion active social media accounts in the world, expanding with an average rate of six new users being created every single second. It's a sharp tool used by Information Warfare actors in malign operations to target beliefs, polarise societies and change behaviours with potential social, political and economic repercussions. In order to effectively counter Information Warfare, it's therefore crucial to learn how to effectively amplify the protective messages, and convey communication according to the values, beliefs and interests of the audience.

Having that said, social media and digital tools and technologies should become strategically analysed and used by state institutions in their efforts to educate and protect their citizens in the larger context of the ever-present Information Warfare. It remains debatable *how* can states become more effective in protecting audiences using media literacy and education, starting from a younger age, involving not only state institutions, but also social actors. We will investigate these open questions further on, in a larger PhD research paper that focuses on consolidating a society's security culture through better institutional strategic communication.

References

Bulao, J., 2023. *How much data is created in 2023.* [Online] Available at: https://techjury.net/blog/how-much-data-is-created-every-day/#gref [Accessed 12 January 2023].

Datareportal, 2022. *Global Social Media Statistics.* [Online] Available at: https://datareportal.com/social-media-users [Accessed 22 December 2022].

Romano, A., 2018. *The Facebook data breach wasn't a hack. It was a wake-up call.* [Online] Available at: https://www.vox.com/2018/3/20/17138756/facebook-data-breach-cambridge-analytica-explained [Accessed 14 January 2023].

Tzu, S., 2011. The art of war. In: s.l.:HarperCollins Publishers.

Bailyn, B., 1992. *The Ideological Origins of the American Revolution.* [Online] Available at: https://military-history.fandom.com/wiki/Winning_hearts_and_minds [Accessed 15 january 2023].

Bao P., H. U., 2020. *An analytical model for the strategy of winning hearts and minds.* [Online] Available at: https://cradpdf.drdc-rddc.gc.ca/PDFS/unc346/p812125_A1b.pdf

Stupples, D., 2015. *What is Information Warfare?.* [Online] Available at: https://www.weforum.org/agenda/2015/12/what-is-information-warfare/ [Accessed 19 December 2022].

Severinghaus, R., 2019. *Seven-Domain Warfare: What Would Mahan Think?.* [Online] Available at: https://blog.usni.org/posts/2019/06/25/seven-domain-warfare-what-would-mahan-think [Accessed 17 January 2023].

Lesenciuc, A., 2016. Razboiul informational. In: Brașov: "Henri Coanda" Air Force Academy Publishing House, p. 98.

Lewis, B. C., 2022. *Information Warfare.* [Online] Available at: https://irp.fas.org/eprint/snyder/infowarfare.htm [Accessed 13 December 2022].

Ullman, H., 2021. *Al cincilea cavaler al apocalipsei și noul MAD.* Bucharest: Ed Militară.

Lucas, E., 2014. *The New Cold War.* London: Bloomsbury.

Scutaru, G., 2022. *Russian Policy in the Black Sea region Report,* Bucharest: New Strategy Center.

Letzing, J., 2022. *John Letzing.* [Online] Available at: https://www.weforum.org/agenda/2022/04/what-is-information-warfare-and-how-pervasive-is-it/ [Accessed 5 January 2023].

LeBon, G., 2022. *Psihologia multimilor,.* Bucharest: Librex,.

Hofstede, G., 2006. *What about Romania.* [Online] Available at: https://www.hofstede-insights.com/country/romania/ [Accessed 13 December 2022].

David, D., 2015. *Psihologia poporului roman.* Bucuresti: Polirom.

Antena3, 2022. *Antena 3.* [Online] Available at: https://www.antena3.ro/emisiuni/news-hour-with-cnn/romania-criza-ulei-floarea-soarelui-ministrul-agriculturii-638177.html [Accessed 22 December 2022].

EuropaLibera, 2022. *EuropaLibera.* [Online] Available at: https://romania.europalibera.org/a/de-unde-a-pornit-panica-cresterii-preturilor-la-benzina-si-motorina/31745801.html [Accessed 20 December 2022].

Munteanu, D., 2022. *Barometrul rezilientei societale,* Bucuresti: Centrul de rezilienta Euroatlantica.

RomaniaLibera, 2022. *Romania Libera.* [Online] Available at: https://evz.ro/romanii-isi-retrag-masiv-banii-de-pe-card-se-intampla-in-plin-razboi-la-granita-romaniei-anuntul-bancilor.html [Accessed 20 December 2022].

Government, R., 2023. *Romania's response to the Ukrainian refugee crisis,* Bucufresti: Guvernul Romaniei.

GlobalFocus, 2023. *Disinformation in a regional context during the War in Ukraine. A comparison between Hungary, Poland, Romania and Serbia,* Bucuresti: Global Focus.

G4Media, 2022. *G4 Media.* [Online] Available at: https://www.g4media.ro/mapn-atentioneaza-asupra-fake-news-urilor-privind-ordine-de-inrolare-in-armata-si-introducerea-serviciului-militar-obligatoriu-va-rugam-sa-va-informati-doar-din-sursele-oficiale.html [Accessed 28 December 2022].

Inforadar, R., 2023. *Inforadar.* [Online] Available at: https://inforadar.mapn.ro/tema28/index.php [Accessed 13 January 2023].

Pelletier, J., 2022. Intelligence, Information Warfare, cyber warfare, electronic warfare – what they are and how Russia is using them in Ukraine. *The Conversation : Science + Technology; Boston* , 1 March.

GlobalFocus, 2023. *Global Focus.* [Online] Available at: https://www.global-focus.eu/2023/01/monitoring-report-on-the-evolution-of-the-main-pro-kremlin-voices-on-facebook-early-warning/ [Accessed 14 January 2023].

Preda, A., 2021. *Social media in Information Warfare. Assault weapon with high recoil.* [Online] Available at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwiGkPTvxs_8AhVJgosKHRZJBd0QFnoECCAQAQ&url=https://revista.unap.ro/index.php/XXI_FSA/article/d [Accessed 7 January 2023].

FM3-05.30, n.d.

FM3-05.302, 2005. *Tactical Psychological Operations Tactics, Techniques, and Procedures.* Washington: US Departament of Army.

FM3-05.302, 2005. s.l.:s.n.

Technology-Oriented Innovations and Cyber Security Challenges in the Healthcare Delivery System: A Perspective from a Developing Economy

Victor Kwarteng Owusu and Ales Gregar
Tomas Bata University in Zlin, Czech Republic

owusu@utb.cz
gregar@utb.cz

Abstract: There is no dispute about the looming digital transformation of certain sectors within transitional economies, especially in Ghana. In fact, for most developed economies, digitalization has proven to have relevant visible effects. The paradox, however, is the seamless nature of this propagation in contrast to the myriad associated benefits. Suffice it to say that, the main purpose of this study was to identify the perceived impediments against the adoption and continuous acceptance of technology-oriented innovation for healthcare delivery in transitional economies through innovations. Especially, in the current technological dispensation where Africa and other developing countries are striving to bridge the technology gap in service delivery. Though, much research has been conducted within the healthcare sector, factors that hinder technology acceptance and continuous usage have rarely been their primary focus. To address this consequential lacuna, this study takes inspiration from literature through document and theme analysis and proposes a model which could serve as a remedy to the identifiable impediments and inhibitors to the seamless delivery of healthcare services in transitional economies. We emphasize good management and government-based interventions throughout the literature review as well as the document analyzed. We also made suggestions for further research, notably concerning means of increasing technology diffusion and possible remedy to cyber-security threats in the healthcare sector.

Keywords: Digital Innovations, Transitional Economies, Cyber-Security, Technology-Oriented and Healthcare Delivery.

1. Introduction

In as many ways imaginable, technological advancement has impacted individuals' work and personal lives in various ways for several decades (Jain & Ranjan, 2020). The current generation is reliant on an era of an unparalleled digital revolution in which firms and organizations rely on the exponential increase in performance through technological leveraging at least in simple terms (Allam, 2019; Costa & Matias, 2020). The way individuals live, communicate, and work has been revolutionized by the evolution of technology in the form of industry 4.0's digital transformation of the business world through technologies such as artificial intelligence (AI), 3D printing and cloud computing. Technologies such as smartwatches, smart homes and online food delivery in the consumer space have also gone through a revolution (Allam, 2019). Again, firms' ability to steer through a turbulent business environment with exponential growth over the years is enhanced by these rapidly emerging information and communication technologies (ICTs) and technology-oriented innovations (Costa & Matias, 2020). Moreover, value creation in the day-to-day management of organizations is significantly influenced by the internet of things (IoT)(Bhatt & Chakraborty, 2021). Moreover, a key strategic benefit for the creation of knowledge by firms is facilitated by big data (Shamim et al., 2019). Unquestionably, lucrative job opportunities have contributed to the improvement of the quality of individual lives due to the favourable business environment of growth created by technology-oriented innovations (Bednárik, 2019).

The healthcare sector in particular has experienced an influx tidal wave of technological innovations for the past few years (Saracci, 2018). The healthcare sector has not only undergone a metamorphosis in multiple digitally-driven technological transformations in diagnostics and equipment but has also experienced an improvement in the administration, management and delivery (Kornyo, 2021). In spite, of enormous technology-driven benefits, such as a decrease in diagnostic errors, and improvement in healthcare delivery, yet, there is seeming resistance to technology by patients and clients alike to these innovations and advancement in digital healthcare technologies in transitional economies (Hira et al., 2022). Therefore, the main purpose of this study is to examine those possible factors that serve as hindrances and inhibitors to professionals in the healthcare sector towards e-health innovations that have proven even more useful in the mix of the pandemic. A conceptual framework based on the qualitative method is advanced subsequently to establish the link between technology-driven innovation and continuous adoption by both healthcare employees and patients. Managerial contributions of this study are to aid policymakers within the healthcare sector in developing a counterargument to condition industry players to accept technology-oriented innovations. Again, the scope of technology acceptance of innovative-driven technologies by stakeholders in the healthcare sector will be broadened in the literature.

The paper subsequently continues with a review of the theoretical foundation of technology-oriented innovations in the healthcare industry in transitional economies. A research methodology which specifies the qualitative process is also elaborated. Lastly, research inferences, deductions, limits and imminent exploration guidance are well-explicated subsequently.

2. Literature Review and Research Propositions

2.1 Infrastructural and Resource constraints

According to Chitungo et al., (2021), the term resource constraints are adopted to indicate the inadequate resources available in developing countries, for instance, the lack of adequate devices and internet connectivity hinders the use of e-health innovations. Different studies have expressed the consequences associated with these resource challenges affecting the operationalization of innovations in developing countries (Clohessy & Acton, 2019). In general, limited access to information through applications, the inability of aged and other vulnerable groups, lack of accessible and compatible mobile phones as well as patients', access to phones that are not supportive of downloads of applications or enabling live consultation (Banskota et al., 2020). Amongst these challenges, are the fact that poorer families' inability to access the internet presupposes that they must rely on the data plan by their network providers which ultimately will increase their expenses and diminish their standard of living should they choose the option of engaging in virtual health session (Van Deursen & Mossberger, 2018). In addition, the quality of software, the limited availability of hardware, as well as the planning and implementation of the overall digitization system depict infrastructural obstacles in healthcare delivery and e-health in developing countries (Kornyo, 2021). Lack of access to computers and internet sources in developing countries, especially, in the hinterlands has become an impediment to healthcare delivery. Whereas, computers and other required gadgets are needed in every facet of the healthcare delivery chain, yet, mostly, only one practitioner can access a computer at a time. Consequently, the primary issue of lack of equipment, which hinders simultaneous use, stagnates the healthcare delivery process. Moreover, the need for regular server and software upgrades, lack of cross-compatibility, uniformity and organizational level integration which poses challenges for accessing information adds to the complications associated with resources and infrastructure (Allam, 2019; Costa & Matias, 2020; Jain & Ranjan, 2020). Moreover, the difficulty of patients' experiences in the use of e-health innovations. For the diffusion of healthcare innovations, the existence of patient care, and patients' acceptance of e-health innovations are vital. Nonetheless, practical usability issues and technology anxieties have become a hindrance to the adoption and subsequent use of e-health innovations (Shaygan & Daim, 2021). It is an open secret that the key infrastructure for health services in every country is telecommunication and internet access (Viganò et al., 2020). However, the most common means of communication and access to the internet in developing countries is through internet café (Bhatt & Chakraborty, 2021). Even though internet penetration amongst Ghanaians has increased within the last decade, the boldness of both the policy formulation, implementation and acceptance of technology innovations by end-users has a hard nut to crack due to the perceived consequences associated with technology adoption and acceptance (Lindemulder, 2021). Healthcare records and transmission of healthcare data derived from these innovations are strengthened and aided by powerful instruments such as information and communication technologies (Aceto et al., 2020). We, therefore, propose that;

P^1: *infrastructure and resource constraints will significantly increase cyber-security challenges*

2.2 Cyber Security Challenges

As the healthcare sector fast adopts technology for healthcare delivery, cyber security is becoming more important (Nifakos et al., 2021). The basic question today in the healthcare industry now is not about the capacity to invest in cybersecurity, but the amount of investment required to invest in cyber level security to ensure safety for not only their data but their entire network operation (Coventry & Branley, 2018; Demirkan et al., 2020). Cybersecurity attacks have become a common phenomenon it is no longer 'if' it happens it is now 'when' will it happen especially, in developing countries (Demirkan et al., 2020; Michael et al., 2019). According to Demirkan et al., (2020), many theories abound, for example; theories of planned behaviour and protection motivation. Nonetheless, all these theories focus on extrinsic rather than intrinsic factors which deal with matters like individuals' mindset of personal and professional lives explained by decision styles. So, many institutions especially, in developing countries are not adequately prepared to remedy the consequences of cyber-attack (Qasaimeh et al., 2022). However, to defend themselves against cyber criminals, institutions must be more prepared and conversant with cyber knowledge (Diogenes & Ozkaya, 2019). Again, cybercrimes are the

motivation for people looking for personal gains or profit (Nurse, 2018). As a result, sensitive documents can be stolen to make money from competitors within the healthcare sector (Demirkan et al., 2020). Therefore, institutions must research and implement robust security systems to minimize some of these cyber threats and fraud (Mugarura & Ssali, 2020). Safe to say that, there are many ways in which cyber-attacks can be carried out, which many institutions are facing today (Tao et al., 2019). Some of these threats are malware which is used by attackers to disrupt and disable computer systems (Alladi et al., 2020). Most of this malware is delivered via e-mail which is hidden from the employees as something else when they are opening it (Demirkan et al., 2020). Another of these threats is Phishing in which the attackers mirror the layout and communications of trusted businesses and common businesses they interrupt (Alawida et al., 2022). The aim is to obtain extra sensitive data from the institution. Again, ransomware is another cyber security threat, this threat is the most prevalent and rapidly emerging threat in which the attackers try to shut down servers and hold hostage data and files until a ransom demand is adhered to (Bhatt & Chakraborty, 2021). Finally, human errors are one of the threats that aid cyber-attacks for instance, unintended disclosures accidental data deletion or improper disposal of sensitive data open the floodgate for cybercriminals to take advantage of the cyber systems of institutions (Sidhu, 2018). As explained by Aldawood & Skinner, (2019), these challenges can be resolved by cyber-security training and awareness for employees. Overall, healthcare institutions must invest and make sure to protect themselves from these cybersecurity challenges (Bhuyan et al., 2020). We, therefore, propose that;

P^2: *Cyber-security challenges will have a significant effect on the adoption and usage of technology-oriented innovations in developing economies.*

2.3 Resistance to Technology-Oriented Innovations

Resistance to technology-oriented innovations can be defined as a complementary mechanism required to improve service delivery within organizations (Hansen et al., 2019). Therefore, value creation, value proposition and customer models of institutions and their branches are affected by digital transformation (Payne et al., 2021). Safe to say that, this scenario is not peculiar to the healthcare industry. Healthcare quality and cost reduction are enhanced by the huge potential of digital technologies (Dash et al., 2019). Nonetheless, digital technology-oriented innovations suffer resistance, especially in transitional economies. The main attributes of the healthcare sector are its complex, regulated and fragmented nature (Jensen et al., 2022). This is partly because healthcare performance is aided by digital technologies through manifold directions, be it new business models permit, or the reshaping of patients, and healthcare provider relationships (Ciasullo et al., 2022). Consequently, new ways and tools for the detection, diagnosis, the transformation of the entire process of and around healthcare as well as the therapy of diseases are provided (Dash et al., 2019; Payne et al., 2021). Eventually, the entire healthcare sector might be plausibly transformed by innovative and disruptive technologies with remarkable consequences for providers, patients and stakeholders (Emilsson et al., 2020). Moreover, hospitals struggle with the massive changes that occur around them, detected by real-world intuitions (Izquierdo et al., 2021). To comply with the latest standards and regulations for technology-oriented innovation to cater for patient security, healthcare institutions must be able to further manage critical technology-oriented innovation systems (Gleiss & Lewandowski, 2022). The difficulty, therefore, is the simultaneous management of existing technology and its continuous exploration (Ciasullo et al., 2022). Consequently, healthcare institutions must reconsider integrating technology-oriented innovations from startups as one of the vital drivers of diffusion of technology in healthcare to serve as a remedy (Behne, 2021). Even though, the healthcare sector has seen a tremendous transformation in terms of progress in the digital space (Wang et al., 2018). However, some barriers promote resistance to the usage of technology within the healthcare industry (Norris et al., 2019). First, technological constraints ranging from the availability of information communication technology (ICT) in a country to the functionality of the provision of service put impediments on the usage of technology (Aceto et al., 2020). Despite, the Ghana government's effort to work on these infrastructures, yet, the issue of interoperability among new and heritage systems has become problematic (Thiel, 2020). For instance, it is difficult to connect the core systems of hospitals in transitional economies with different technologies (Norris et al., 2019). This is partly because of low compatibility, especially when the available devices cannot operate offered services. Further, most healthcare centres in developing countries do not have network access. Safe to say that, poor data integrity and quality are associated with a lack of data which breeds low reliability and a poor foundation for superior data-based services (Behne, 2021). Processes and structures resulting from workflow deficiencies lead to concerned user groups being excluded from the defect in the decision-making process by these deficiencies (Grim et al., 2019). Again, physicians do not have enough time for non-patient-related issues (Gleiss & Lewandowski, 2022). Another constraint is the attitude towards technology devoid of intrinsic motivation and knowledge (Wang et al., 2018). The negative

response towards technology is a result of a lack of incentives and perceived usefulness and confidence in technology generally, especially in the case of the aged (Gleiss & Lewandowski, 2022). Mistrust towards newly technology-oriented innovations and missing acceptance of digital health acceptance especially, wrong or missing information are all factors that lead to resistance to technology-oriented innovations (Nymberg et al., 2019). Fear of transparency within the medical process is also another constraint that results in a lack of control of clinicians and medical staff while strengthening patients' positions (Shammi et al., 2020). Verification issues and missing public funds are some of the constraints are some of the monetary problems associated with technology-oriented innovations (Gleiss & Lewandowski, 2022). Therefore, we propose that;

***P³**: lack of infrastructure and resources will affect the adoption and usage of technology-oriented innovations in developing economies.*

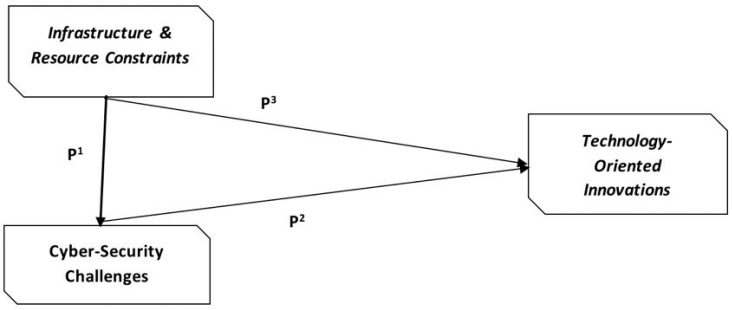


Figure 1: Conceptual Framework

***Source:** Authors own construct depicting the various propositions made.*

3. Methodology

The researchers adopted document analysis as the fulcrum in reaching their overall goals. Safe to say that this was wholly qualitative-based research. Much as document analysis is a first-rate preliminary subject for drawing upon germane phenomena given its wider coverage area, it might be taken for an inexperienced research method. Nonetheless, it is deep-rooted in systematic documents obtained from both paid and open databases such as EBSCO, SCOPUS, Thompsons Reuters WoS, Google scholar and other web-based sources aided by keyword search. Moreover, theme analysis was adopted to identify the relations among the research construct. Theme analysis offered the opportunity in searching for likely links and patterns across numerous research domains taking cognizance of the deductive research aim (Eisenhardt, 2021). Subject to mined information, the secondary sources tied with the researchers' knowledge, the related domains identified were linked to the main research constructs, technology-oriented innovations. In addition, due to the array of literature dealing with the broader theme of technology-oriented innovations. Document analysis is an appropriate method since it lays bare the copious background to studying the variation of the study's phenomenon. The themes of the study were usually as the case may be gathered from the organizational performance strategy and particularly from research-based concepts (RBC) and the research domain of technology-oriented innovations and the continuous usage within the healthcare sector. Overall, the researchers have more or less organized as many themes of the relevant content through document analysis, to accomplish the desired objectives of the study. This concept was aimed at building a conceptual model. (**see Figure 1**). Readers are implored to examine the works of others for the advocacy of document analysis as a treasured method for conducting a qualitative analysis such as the current study. Again readers are encouraged to find the works of Khitous et al., (2020), and Best et al., (2017), (2019) for a dynamic read on themes analysis. The researchers unequivocally do not claim the thorough documentary analysis and the resultant themes that arose from the procedure. Nonetheless, an effort has been made to expound critically related themes of the research to extend the literature on the phenomenon under study.

4. Discussions, Implications and Research Limitations

A postmortem analysis of other studies concerning technology-oriented innovations characterizes this work. Primarily, the researchers try to obtain empirical data from a transitional economic country to analyze the reluctance of both healthcare employees and patients to accept technology-oriented innovations in the Ghanaian healthcare space. As stated by Saracci, (2018), the healthcare industry has experienced a tidal wave

of technological innovations for the past few decades. Consequently, the healthcare sector has gradually metamorphosed into multiple digitally-driven technological transformations in diagnostics and equipment but has also experienced an improvement in administration as suggested by Kornyo, (2021). According to Hira et al., (2022), even though, there have been enormous benefits derived from technology-driven innovations, such as a decrease in diagnostic errors, and improvement in healthcare delivery, yet, there is seemingly resistance to these innovations and advancements of technologies in developing economies. This situation is primarily because of circumstances such as infrastructural and resource constraints, as well as cyber-security challenges.

First, insufficient resources such as device and internet connectivity and availability in developing countries hinder the application of technology-driven innovations as explained by Chitungo et al., (2021). Banskota et al., (2020), believe that, generally, inhibitors such as limited access to information through applications, the inability of the aged and other vulnerable groups to apply these gadgets, accessibility of compatible mobile phones and access to mobile phones which are unable to download an application or unable for live consultations, especially in the hinterlands. Moreover, poorer families' inability to access the internet indicates that they must rely on a data plan by network providers which ultimately increase their overall expenses and diminish their standard of living, should they choose the option of engaging in virtual health session (Van Deursen & Mossberger, 2018). The quality of software, the limited availability of hardware, and the planning and implementation of the overall digitization system is an indication of infrastructure obstacles in healthcare delivery in developing countries (Clohessy & Acton, 2019). As long as these technological devices are scarce, some of these hindrances will continue to exist. Even worse, is the fear of cyberattacks by unscrupulous people who will do anything to get hold of sensitive data to the detriment of patients and employees of the healthcare sector for their self-parochial interest. According to Demirkan et al., (2020), cyberattacks are no longer 'if it happens' it is now about 'when will it happen?'. The basic question then is not about the capacity to invest in cyber-security but the magnitude of investment that will ensure safety for not only data but network operations (Coventry & Branley, 2018).

According to Kouhizadeh et al., (2021), organizational constraints are factors that hinder the exploration of technology-oriented innovations in organizational settings and culture within the healthcare delivery system. As observed by Carney, (2011); & Duffy, (2018) Healthcare organizations are driven by care, healing and quality of service, for healthcare organizations (Abuosi et al., 2022), to strive for excellence in meeting the needs of their patients, through a focused and safety environment (Gleiss & Lewandowski, 2022). Globally, there is a need to develop a culture of safety where adverse measures are reported, analyzed and critiqued. After all, there is a connection between a healthy workplace environment and a healthy patient (Munro & Hope, 2020). In addition, as stated by Curtis et al., (2019) needed competencies for the achievement of high reliability on the assurance of safe patient care are based on team and technical competence. Again, for healthcare institutions to disseminate technology-oriented innovations in developing countries, clinicians and non-clinicians must be strategic thinkers and also understand, know and appreciate the mission for a clear and focused strategic direction (Aufegger et al., 2020).

Overall, time and effort must be invested in top-level managers, clinicians and non-clinicians to know and understand the organizational strategy to improve the strategic thinking which buttresses strategic planning at the top level (Aufegger et al., 2020). Moreover, according to Payne et al., (2021), value creation, value propositions and customer models of institutions and their branches are affected by digital information and technology-oriented innovations. Even though healthcare quality and cost reduction are enhanced by the huge potential of technology-oriented innovations, it suffers resistance, especially in transitional economies. The Healthcare industry has an attribute of fragmented, complex and regulatory issues (Hansen et al., 2019). This is partly because healthcare performances are aided by digital technologies via manifold directions, be it new business model permits, or the reshaping of the relationship between patients and healthcare providers as explained by Ciasullo et al., (2022). Consequently, new ways and tools for the detection, diagnosis and transformation of the entire process of and around healthcare as well as therapy of diseases are provided. Nonetheless, hospitals struggle with massive changes that occur around detected by real-world intuition (Izquierdo et al., 2021). To comply with the latest standards and regulations for technology-orientated innovations and patient security, healthcare institutions have to be able to further manage critical technology-oriented innovation systems (Gleiss & Lewandowski, 2022). The difficulty, therefore, is simultaneous management of the existing technology, information and its exploration, consequently, healthcare institutions must reconsider integrating information technology-oriented innovation. We can therefore say that all the propositions (P^1, P^2 & P^3) are supported.

4.1 Implications of The Study

4.1.1 Theoretical, Managerial and Practical Implications.

The focal point of this research is the dissection of factors that hinders the acceptance of technology-oriented innovations by the healthcare industry. Even though there are enormous benefits of technology-driven innovations there is seeming resistance to these innovations and advancement in digital healthcare technologies in developing countries. Therefore, the main purpose of this study is to examine factors which inhibit the adoption of technology-oriented innovations which has proven to be even more useful in the mix of covid-19 within the healthcare industry in developing countries. This study provides useful insights for healthcare managers in their policy formulation and implementation. Generally, healthcare facilities invest in physical infrastructure as opposed to technology infrastructure. This is primarily may be due to the resistance of patients and clients to technology-oriented innovations usage (Gleiss & Lewandowski, 2022). Nonetheless, this study recommends a tangible indication of how investment in technology-oriented innovations will promote a proper healthcare delivery system. The study also looks at how an existent literature review gives academics theoretical insights into some of the constraints that hinder the adoption of technology-oriented innovations. Most qualitative research on the aforementioned topic is conducted in developed economies. Therefore, our study seeks to add a significant contribution to the of adoption technology-oriented innovations for the good of the healthcare industry in developing countries. An orientational guide in aiding technology-oriented innovations in policy formulation and planning is the focal point of this study. Naturally, resources in technology are considered as a cost rather than an investment due to the resistance from an organizational point of view. Because technology-related investments fail to yield actual tangible returns. Moreover, whereas organizations prefer the investment of more resources into visible infrastructure in their policy formulation other than technology, especially in developing countries, yet, the current trend calls for the adoption and the introduction of technology-oriented innovations other than physical infrastructure. This study provides a persuasive reason why the healthcare industry must invest in technology-driven innovations. Due to the resistance to the adoption of technology-oriented innovations, lack of infrastructure and other resourced constraints. For instance, the Ghana government through digitalization policy is trying to invest in aspects of these technology-driven innovations to remedy the aforementioned obstacles. However, there is still the challenge of technology usage amongst the people especially those in the hinterlands. Therefore, showcase the bond between the availability of digital technology amongst the citizenry and the policy and planning of the management of the various healthcare facilities. For instance, when managers of health facilities invest in infrastructure and technology-oriented innovations, it stands to improve health delivery but, the problem remains the usability by clients and patients. Suffice it to say that; data and network security are equally relevant as cyberattack security has become a common phenomenon. Therefore, for both patients and health professionals to adopt technology-oriented innovations, managers in the healthcare sector must be ready to invest in robust security systems to minimize cyber threats and fraud.

5. Research Limitations and Future Research Directions

Our study not only supplements accumulated literature with valuable insights but also proposes real-world contributions. In contrast, notwithstanding the substantial contributions, as with all other academic enquiries, our study involves certain limitations in its method and scope. A crucial prerequisite for acknowledging future research directions is susceptible to these limitations. Regarding the methodological limitations, a wholly conceptual study via qualitative research was adopted for this enquiry. Future research can adopt a different research method like mixed methodology or a wholly empirical method to yield more robust insights. Regarding the scope, firstly, the technology-driven innovations as described in this work stem from the works of (Aufegger et al., 2020; Banskota et al., 2020; Carney, 2011; Chitungo et al., 2021; Demirkan et al., 2020; Gleiss & Lewandowski, 2022; Qasaimeh et al., 2022). Safe to admit that, empirically validated data will be enough to affirm or prove otherwise, despite, the reluctance of some organizations to release data due to concerns for security perceived to be sensitive. There are many parts of qualitative research, yet, our work is wholly centred on a conceptual model, therefore, future studies may consider other qualitative models like surveys or in-depth interviews to broaden the horizon of the adoption of information and technology-oriented innovations to make the healthcare delivery effective and efficient.

References

Abuosi, A. A., Poku, C. A., Attafuah, P. Y. A., Anaba, E. A., Abor, P. A., Setordji, A., & Nketiah-Amponsah, E. (2022). Safety culture and adverse event reporting in Ghanaian healthcare facilities: Implications for patient safety. *Plos One*, *17*(10), e0275606.

Aceto, G., Persico, V., & Pescapé, A. (2020). Industry 4.0 and health: Internet of things, big data, and cloud computing for healthcare 4.0. *Journal of Industrial Information Integration*, *18*, 100129.

Alawida, M., Omolara, A. E., Abiodun, O. I., & Al-Rajab, M. (2022). A deeper look into cybersecurity issues in the wake of Covid-19: A survey. *Journal of King Saud University-Computer and Information Sciences*.

Aldawood, H., & Skinner, G. (2019). Reviewing cyber security social engineering training and awareness programs—Pitfalls and ongoing issues. *Future Internet*, *11*(3), 73.

Alladi, T., Chamola, V., & Zeadally, S. (2020). Industrial control systems: Cyberattack trends and countermeasures. *Computer Communications*, *155*, 1–8.

Allam, Z. (2019). *Cities and the Digital Revolution: Aligning technology and humanity*. Springer.

Amran, A., Ananda, A., & Khairani, S. (2018). Effectiveness of integrated science instructional material on the pressure in daily life theme to improve digital age literacy of students. *Journal of Physics: Conference Series*, *1006*(1), 12031.

Aufegger, L., Alabi, M., Darzi, A., & Bicknell, C. (2020). Sharing leadership: current attitudes, barriers and needs of clinical and non-clinical managers in UK's integrated care system. *BMJ Leader*, leader-2020.

Banskota, S., Healy, M., & Goldberg, E. M. (2020). 15 smartphone apps for older adults to use while in isolation during the COVID-19 pandemic. *Western Journal of Emergency Medicine*, *21*(3), 514.

Bednárik, J. (2019). Change of paradigm in personnel strategy–corporate social responsibility and internal communication. *Communication Today*, *10*(2), 42–57.

Behne, A. (2021). *Process-driven Innovation: An Analysis of Digital Health Technologies*.

Best, P., Badham, J., Corepal, R., O'Neill, R. F., Tully, M. A., Kee, F., & Hunter, R. F. (2017). Network methods to support user involvement in qualitative data analyses: an introduction to participatory theme elicitation. *Trials*, *18*(1), 1–14.

Best, P., McConnell, T., Davidson, G., Badham, J., & Neill, R. D. (2019). Group-based video-conferencing for adults with depression: findings from user-led qualitative data analysis using participatory theme elicitation. *Research Involvement and Engagement*, *5*(1), 1–11.

Bhatt, V., & Chakraborty, S. (2021). Improving service engagement in healthcare through the internet of things based healthcare systems. *Journal of Science and Technology Policy Management*.

Bhuyan, S. S., Kabir, U. Y., Escareno, J. M., Ector, K., Palakodeti, S., Wyant, D., Kumar, S., Levy, M., Kedia, S., & Dasgupta, D. (2020). Transforming healthcare cybersecurity from reactive to proactive: current status and future recommendations. *Journal of Medical Systems*, *44*(5), 1–9.

Carney, M. (2011). Influence of organizational culture on quality healthcare delivery. *International Journal of Health Care Quality Assurance*.

Chitungo, I., Mhango, M., Mbunge, E., Dzobo, M., Musuka, G., & Dzinamarira, T. (2021). Utility of telemedicine in sub-Saharan Africa during the COVID-19 pandemic. A rapid review. *Human Behavior and Emerging Technologies*, *3*(5), 843–853.

Ciasullo, M. V., Orciuoli, F., Douglas, A., & Palumbo, R. (2022). Putting Health 4.0 at the service of Society 5.0: Exploratory insights from a pilot study. *Socio-Economic Planning Sciences*, *80*, 101163.

Clohessy, T., & Acton, T. (2019). Investigating the influence of organizational factors on blockchain adoption: An innovation theory perspective. *Industrial Management & Data Systems*.

Costa, J., & Matias, J. C. O. (2020). Open innovation 4.0 as an enhancer of sustainable innovation ecosystems. *Sustainability*, *12*(19), 8112.

Coventry, L., & Branley, D. (2018). Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. *Maturitas*, *113*, 48–52.

Curtis, E., Jones, R., Tipene-Leach, D., Walker, C., Loring, B., Paine, S.-J., & Reid, P. (2019). Why cultural safety rather than cultural competency is required to achieve health equity: a literature review and recommended definition. *International Journal for Equity in Health*, *18*(1), 1–17.

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and prospects. *Journal of Big Data*, *6*(1), 1–25.

Demirkan, S., Demirkan, I., & McKee, A. (2020). Blockchain technology in the future of business cyber security and accounting. *Journal of Management Analytics*, *7*(2), 189–208.

Diogenes, Y., & Ozkaya, E. (2019). *Cybersecurity–Attack and defence strategies: Counter modern threats and employ state-of-the-art tools and techniques to protect your organization against cyber criminals*. Packt Publishing Ltd.

Duffy, J. R. (2018). *Quality caring in nursing and health systems: Implications for clinicians, educators, and leaders*. Springer Publishing Company.

Eisenhardt, K. M. (2021). What is the Eisenhardt Method? *Strategic Organization*, *19*(1), 147–160.

Emilsson, M., Ernstson, U., Gustavsson, L., & Svensson, A. (2020). Sustainable innovations in small enterprises for the transformation of the primary healthcare sector. *Sustainability*, *12*(16), 6391.

Gleiss, A., & Lewandowski, S. (2022). Removing barriers to digital health through organizing ambidexterity in hospitals. *Journal of Public Health*, *30*(1), 21–35.

Grim, K., Tistad, M., Schön, U.-K., & Rosenberg, D. (2019). The legitimacy of user knowledge in decision-making processes in mental health care: An analysis of epistemic injustice. *Journal of Psychosocial Rehabilitation and Mental Health*, *6*(2), 157–173.

Hansen, E. G., Wicki, S., & Schaltegger, S. (2019). Structural ambidexterity, transition processes, and integration trade-offs: a longitudinal study of failed exploration. *R&D Management*, *49*(4), 484–508.

Hira, F. A., Khalid, H., Rasid, S. Z. A., Baskaran, S., & Moshiul, A. M. (2022). *Blockchain Technology Implementation for Medical Data Management in Malaysia: Potential, Need and Challenges*.

Izquierdo, J. L., Morena, D., González, Y., Paredero, J. M., Pérez, B., Graziani, D., Gutiérrez, M., & Rodríguez, J. M. (2021). Clinical management of COPD in a real-world setting. A big data analysis. *Archivos de Bronconeumología (English Edition)*, *57*(2), 94–100.

Jain, A., & Ranjan, S. (2020). Implications of emerging technologies on the future of work. *IIMB Management Review*, *32*(4), 448–454.

Jensen, N., Barry, A., & Kelly, A. H. (2022). More-than-national and less-than-global: The biochemical infrastructure of vaccine manufacturing. *Economy and Society*, 1–28.

Khitous, F., Strozzi, F., Urbinati, A., & Alberti, F. (2020). A systematic literature network analysis of existing themes and emerging research trends in a circular economy. *Sustainability*, *12*(4), 1633.

Kornyo, E. A. (2021). *Exploring the use of Artificial Intelligent Systems in STEM Classrooms*. Columbia University.

Kouhizadeh, M., Saberi, S., & Sarkis, J. (2021). Blockchain technology and the sustainable supply chain: Theoretically exploring adoption barriers. *International Journal of Production Economics*, *231*, 107831.

Lindemulder, A. (2021). *Social innovation in developing economies The success versus the failure of social innovation projects in reducing economic inequality in developing economies*.

Michael, K., Kobran, S., Abbas, R., & Hamdoun, S. (2019). Privacy, data rights and cybersecurity: Technology for good in the achievement of sustainable development goals. *2019 IEEE International Symposium on Technology and Society (ISTAS)*, 1–13.

Mugarura, N., & Ssali, E. (2020). Intricacies of anti-money laundering and cyber-crimes regulation in a fluid global system. *Journal of Money Laundering Control*, *24*(1), 10–28.

Munro, C. L., & Hope, A. A. (2020). Healthy work environment: resolutions for 2020. In *American Journal of Critical Care* (Vol. 29, Issue 1, pp. 4–6). AACN.

Nifakos, S., Chandramouli, K., Nikolaou, C. K., Papachristou, P., Koch, S., Panaousis, E., & Bonacina, S. (2021). Influence of human factors on cyber security within healthcare organisations: A systematic review. *Sensors*, *21*(15), 5119.

Norris, J. M., Zhuo, A., Govendir, M., Rowbotham, S. J., Labbate, M., Degeling, C., Gilbert, G. L., Dominey-Howes, D., & Ward, M. P. (2019). Factors influencing the behaviour and perceptions of Australian veterinarians towards antibiotic use and antimicrobial resistance. *PloS One*, *14*(10), e0223534.

Nurse, J. R. C. (2018). Cybercrime and you: How criminals attack and the human factors that they seek to exploit. *ArXiv Preprint ArXiv:1811.06624*.

Nymberg, V. M., Bolmsjö, B. B., Wolff, M., Calling, S., Gerward, S., & Sandberg, M. (2019). 'Having to learn this so late in our lives…'Swedish elderly patients' beliefs, experiences, attitudes and expectations of e-health in primary health care. *Scandinavian Journal of Primary Health Care*, *37*(1), 41–52.

Payne, E. H. M., Peltier, J., & Barger, V. A. (2021). Enhancing the value co-creation process: artificial intelligence and mobile banking service platforms. *Journal of Research in Interactive Marketing*.

Qasaimeh, M., Hammour, R. A., Yassein, M. B., Al-Qassas, R. S., Torralbo, J. A. L., & Lizcano, D. (2022). Advanced security testing using a cyber-attack forecasting model: A case study of financial institutions. *Journal of Software: Evolution and Process*, *34*(11), e2489.

Saracci, R. (2018). Epidemiology in wonderland: Big Data and precision medicine. *European Journal of Epidemiology*, *33*(3), 245–257.

Shamim, S., Zeng, J., Shariq, S. M., & Khan, Z. (2019). Role of big data management in enhancing big data decision-making capability and quality among Chinese firms: A dynamic capabilities view. *Information & Management*, *56*(6), 103135.

Shammi, M., Bodrud-Doza, M., Islam, A. R. M. T., & Rahman, M. M. (2020). COVID-19 pandemic, socioeconomic crisis and human stress in resource-limited settings: a case from Bangladesh. *Heliyon*, *6*(5), e04063.

Shaygan, A., & Daim, T. (2021). Technology management maturity assessment model in healthcare research centres. *Technovation*, 102444.

Sidhu, S. (2018). *Information security in the healthcare system*.

Tao, H., Bhuiyan, M. Z. A., Rahman, M. A., Wang, G., Wang, T., Ahmed, M. M., & Li, J. (2019). Economic perspective analysis of protecting big data security and privacy. *Future Generation Computer Systems*, *98*, 660–671.

Thiel, A. (2020). Biometric identification technologies and the Ghanaian 'data revolution.' *The Journal of Modern African Studies*, *58*(1), 115–136.

Van Deursen, A. J. A. M., & Mossberger, K. (2018). Anything for anyone? A new digital divide in internet-of-things skills. *Policy & Internet*, *10*(2), 122–140.

Viganò, E., Loi, M., & Yaghmaei, E. (2020). Cybersecurity of critical infrastructure. In *The Ethics of Cybersecurity* (pp. 157–177). Springer, Cham.

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, *126*, 3–13.

Deep-learning-based Intrusion Detection for Software-defined Networking Space Systems

Uakomba Uhongora[1], Ronald Mulinde[1], Yee Wei Law[1] and Jill Slay[1,2]
[1]UniSA STEM, University of South Australia, Mawson Lakes, Australia
[2]SmartSat CRC, Adelaide, Australia

uakomba.uhongora@mymail.unisa.edu.au
Ronald.Mulinde@unisa.edu.au
YeeWei.Law@unisa.edu.au
Jill.Slay@unisa.edu.au

Abstract: This paper briefly reviews the application of the Software-defined Networking (SDN) architecture to satellite networks. It highlights the prominent cyber threats that SDN-based satellite networks are vulnerable to and proposes relevant defence mechanisms. SDN transforms traditional networking architectures by separating the control plane from the forwarding (data) plane. This separation enhances scalability and centralises management. In comparison, in traditional networks, the control plane and the data plane are usually combined, resulting in complex network management and reduced scalability. Satellite networks can take advantage of these benefits offered by SDN and this supports them as key enablers of critical services, including weather prediction, global broadband Internet coverage, and Internet of Things (IoT) services. Ease of configuration and flexibility are essential for satellites providing critical services to instantly adapt to network changes. These desirable attributes can be realised by applying SDN to satellite networks. Although SDN offers significant benefits to satellite networks, it is vulnerable to cyber-attacks and particularly due to its centralised architecture. A common attack on SDN is the Distributed Denial of Service (DDoS) attack which could render the entire SDN unavailable. To mitigate such threats, an efficient Intrusion Detection System (IDS) is required to monitor the network and detect any suspicious traffic. However, traditional IDSs produce too many false positives and often fail to detect advanced attacks. For their ability to learn feature hierarchies in network traffic data automatically, whether, for network traffic classification or anomaly detection, deep learning (DL) plays an increasingly important role in IDSs. In this paper, we present a brief review of recent developments in cyber security for SDN-based space systems, and we identify vulnerabilities and threats to an SDN-based satellite network. We further discuss the potential of a DL-based IDS for the detection of cyber threats. Finally, we identify further research gaps in the recent literature and propose future research directions.

Keywords: Software-defined Networking, Space Systems, Satellite Networks, Deep Learning, Intrusion Detection System.

1. Introduction

1.1 Background

A space system is a set of interrelated or interacting components that together meet one or more space-related objectives (Aguirre, 2013); examples of these components include ground control systems, launch facilities, suborbital vehicles, orbital vehicles, and space probes. Figure 1 shows the typical structure of a space system, consisting of a space segment and a ground segment, which communicate with each other via radio frequency (RF) signals (Manulis et al, 2020) and increasingly optical signals. The space segment comprises standalone or networked satellites in orbits (as well as launch vehicles designed to release satellites into orbits). Manulis et al (2020) describe the structure of the space segment which consists of a satellite containing one or more payloads. Equipment designed to perform satellite functions. A bus that houses the payload and the rest of the satellite system. Major satellite systems include telemetry, tracking and command/control (TT&C), command and data processing (C&DH), and attitude determination and control (ADCS). These subsystems are responsible for receiving and processing uplink and downlink signals, validating, decoding and transmitting commands to other subsystems, and controlling satellite stabilization or pointing.

An intelligent space system is the enhancement of space system assets by integrating sensors, actuators and intelligent technologies that can collect, analyse and communicate information to the end user. Intelligent space systems provide critical services including GPS navigation, broadband global Internet coverage, and IoT services. Remote communication between devices and systems, as well as the extension of coverage to areas without Internet and cellular network, are enhanced by intelligent space systems. Space systems are also critical to the military domain for their Command, Control, Communications, Computer, Intelligence, Surveillance, and Reconnaissance (C4ISR) capabilities (Pavur & Martinovic, 2020).

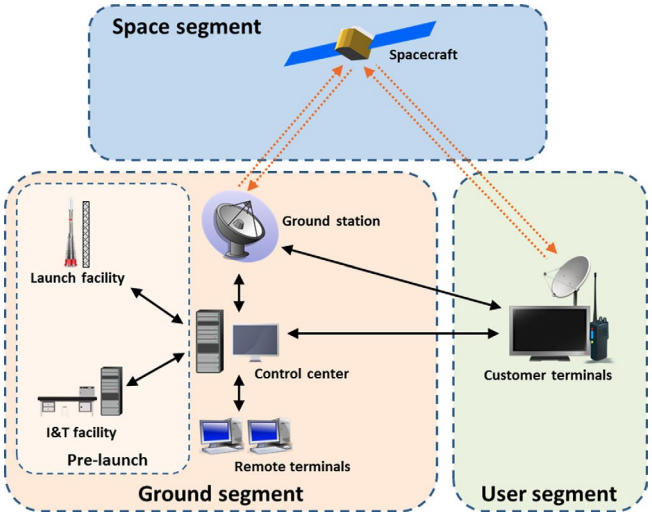


Figure 1: Space system architecture (Manulis et al, 2020)

Unfortunately, space systems are often the target of cyber-attacks (Pavur & Martinovic, 2020). Several critical infrastructures heavily rely on space systems, and these infrastructures would be greatly affected if the space systems are breached. Remote mobile communication around the world would be negatively affected, and people in remote areas that depend on satellites for TV, radio, and the Internet could experience disrupted connectivity. GPS services required by aeroplanes, cargo vessels, and people to navigate geographical locations would be greatly affected too. Moreover, global timing synchronisation would be interrupted. Overall, the absence or disruption of space systems would greatly affect critical domains such as the military, health, critical infrastructure, business and economy, and weather forecast. Considering the dependency of critical infrastructure on space systems, space systems security is required to protect the space ecosystem from cyber threats (Plotnek & Slay, 2022). Space systems security is defined as "the assurance of the confidentiality, integrity, and availability of a space system throughout its lifecycle, including all ground, communications, and space segments as well as the data, processes, and supply chains that support it" (Plotnek & Slay, 2022). Figure 2 illustrates the different threat vectors that can be exploited by hackers on a space system.

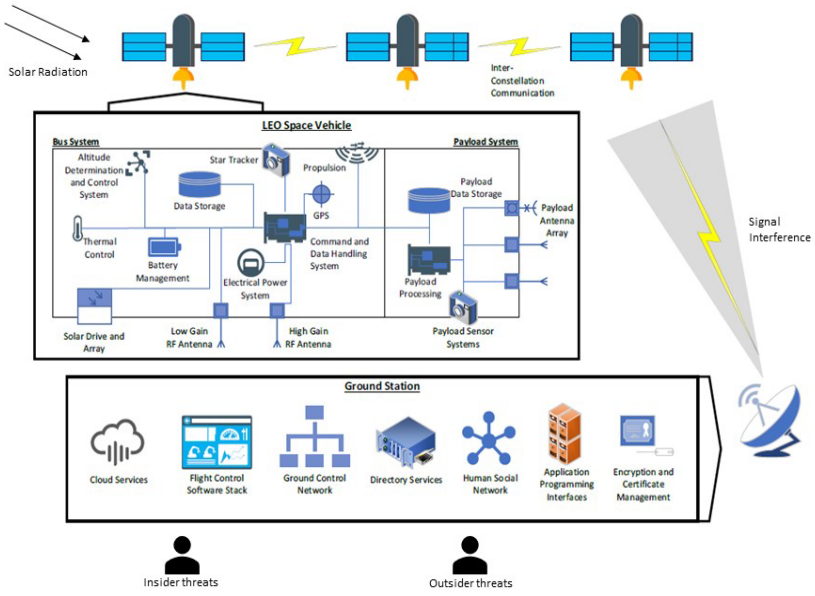


Figure 2: Space system components and threat vectors (Ormrod, Slay & Ormrod, 2021)

Due to the demand for high-performance and worldwide broadband coverage from space systems, space systems can benefit from the high bandwidth offered by SDN at a low cost and efficient packet delivery (Dey and Rahman, 2019). Space systems typically use traditional network design which limits the flexibility and configuration of the network. For space systems to enable services to the critical systems they must be easy to

configure and flexible enough to instantly adapt to network changes. These benefits can be realised by applying SDN to space systems.

SDN is a prominent network architecture that has gained popularity due to its enhancement and improvement to traditional networking. SDN is a networking paradigm that transforms the traditional networking architecture by separating the control plane from the forwarding (data) plane, thus improving the efficiency of applications on how data is transferred by optimising and routing traffic within the network (Pradhan & Mathew, 2020). SDN was developed for the improvement of the shortcomings associated with traditional networks. The controller in the SDN architecture is the "brain" of the network and it decides the paths taken by packets within the network. These networking features allow network administrators and programmers to easily monitor and control the entire network at a limited cost and optimise network resource usage. The SDN architecture enhances scalability, centralises management, and eases the programmability of a network (Bertaux et al, 2015). On the other hand, with traditional networks, the control plane and the data plane are usually combined resulting in complex network management and reduced scalability. Figure 3 illustrates the different architectural designs of SDN in contrast to traditional networks.

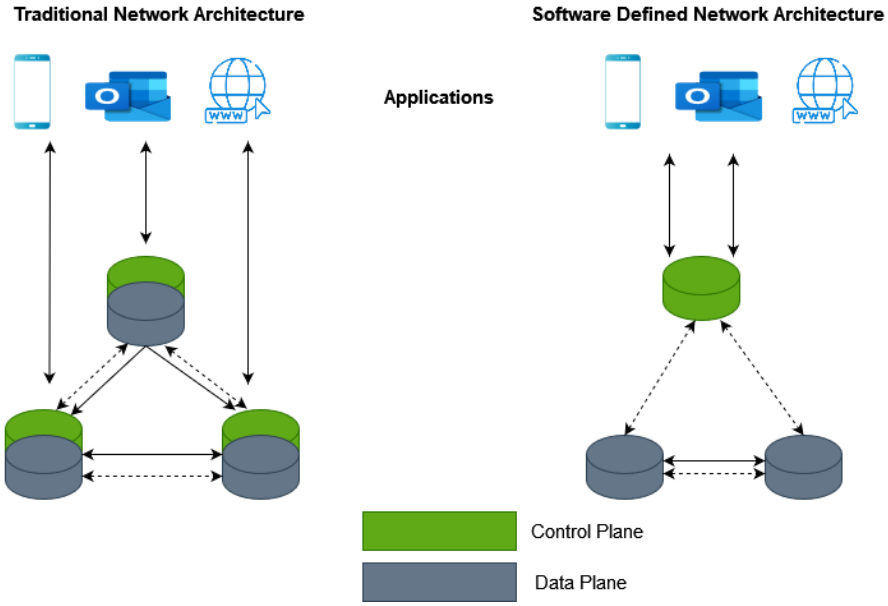


Figure 3: Traditional network architecture vs SDN architecture

Although SDN offers exceptional features for space systems, it brings new security challenges that are unique to SDN (Akhunzada, Ahmed, Gani, Khan and Imran, 2015). The decoupling of the data plane from the control plane in SDN makes it vulnerable to attacks mainly at its control plane. Since the controller makes all the decisions in the network when the controller is compromised the entire network becomes vulnerable. The Distributed Denial of Service (DDoS) attack is one of the most common attacks against the SDN controller. A DDoS attack could flood the controller with large traffic resulting in the unavailability of the network. Network programmability enabled by SDN is a double-edged sword because an attacker could hijack the controller and maliciously reprogram the entire network.

These cyber-attacks amongst others, aim to manipulate data, exhaust network resources and sometimes deny legitimate users access to required network servers. Unfortunately, the adoption of SDN in space systems inherently exposes them to cyber-attacks targeted at SDN. If the vulnerabilities are not managed, SDN-based space systems are exposed to attacks that could render them unavailable or attacks that intercept satellite communication. Thus, there is a need to address the security challenges in SDN-based space systems.

1.2 Research Context

The proposed research aims to address the security issues/cyber threats faced by SDN-based space systems by developing intelligent security controls.
There is a demand for space systems due to their use in a variety of critical systems such as the military, weather services, and global timing. Due to the reliance on critical systems on the space system, cybercriminals see space

systems as a "single point of failure in many critical services" (Pavur & Martinovic, 2020). Space systems experience the challenge of inflexible network management due to their use of traditional networks (Li et al, 2016) that fail to adapt to the demands in bandwidth, computing and storage of current applications. To resolve this problem, adopting SDN for space systems can greatly provide scalability and flexible networking architecture required by critical systems such as space systems. SDN is one of the most promising evolving network technologies (Liu *et al.*, 2020) as it transforms the traditional network architecture into an independent setup where the control plane is separated from the data plane, allowing greater control and programmability functions to the network administrator (Malik *et al.*, 2020).

1.2.1 Software-defined Networking

The existing architectures of space systems are inflexible due to their dependence on hardware (Bertaux et al, 2015). SDN brings flexibility, automation, and customization to space systems (Ferrús et al, 2016; Bertaux et al, 2015). Recent literature has introduced the concept of applying SDN to space systems to benefit from the features of SDN. SDN enables arbitration of resources, e.g., a ground station operator to dynamically switch mission operators to their allocated antenna system at the scheduled time (Liu et al, 2020). SDN provides effective traffic scheduling for space systems (Li et al, 2016). In addition, the SDN-based space system "architecture integrates functions such as data routing, resource allocation, network failure and security monitoring in the satellite network into the controller under the concept of SDN" (Liu et al, 2018). However, these great benefits of SDN come with security vulnerabilities. For instance, the SDN control plane poses a great vulnerability because it is the point of configuration and administration of the entire network and can be a single point of failure during an attack. Unfortunately, SDN vulnerabilities are inevitably transferred to space systems. These vulnerabilities expose space systems to the disruption of network services or accessibility and eavesdropping of signals transmitted between space systems and ground stations (Shadbolt, 2021). There is a need to secure SDN-based space systems, for they enable the connection and efficiency of critical systems. Consequently, we propose intelligent security control that will protect space systems from known and unknown malicious activities.

1.2.2 Cyber threat landscape

Cybercriminals have improved their attack strategies, and this has produced novel attacks that traditional security controls are unable to withstand (Kaloudi & Li, 2020). Cybercriminals have started to integrate Artificial Intelligence (AI) techniques to enhance the impact and ease of performing more powerful and extensive cyber-attacks (Kaloudi & Li, 2020). The use of AI to perform cyber-attacks has bred intelligent attacks that require intelligent security controls. Additionally, the cyber threats faced by space systems originate from electronic and digital sources and kinetic and physical sources and from inside or outside the space system. Thus, the cyber threat landscape for space systems has broadened, leaving the space system exposed to cyber-attacks.

1.3 Research objective and Research questions

The main objective of this research is to develop intelligent security controls for integrated SDN-based space systems. This research aims to answer one major question, which is:

1. *How can intelligent security controls for integrated SDN-based space systems be developed and implemented?*

This research question is supported by a series of ancillary sub-questions:

1. *How is space system security defined and mapped?*
2. *What are intelligent space systems?*
3. *What are SDN vulnerabilities and how are they mitigated?*
4. *How is SDN used for the flexible integration of space and terrestrial systems?*
5. *What kind of attacks is SDN subject to?*

2. Literature review

In this section, we present a review of previous work in our multi-disciplinary domain. Firstly, we analyse work that has examined space systems in context and the cyber threats they face. Secondly, we discuss literature that has focused on SDN, its benefits in contrast to traditional networks, and its vulnerabilities. Thirdly, we present a review of previous work on the cyber threat landscape. It is by reviewing this literature that we identified the

research gap to be addressed by this research. We also explore the potential use of deep-learning-based intrusion detection solutions for our research context and research objective.

2.1 Space systems

2.1.1 Space systems context

Space systems play a major role in the digital communication era. There is an increased demand for space systems to enhance "ubiquitous broadband service and remote sensing capacity" (Pavur & Martinovic, 2020). Fortunately, the procurement and launching of space systems have become so inexpensive that even new organizations in the space industry can afford them.

2.1.2 Space systems security domain

Any component of a space system is a potential attack vector (Falco, 2018). Space systems are vulnerable to cyber-attacks such as hijacking, DDoS, and ransomware.

2.2 Software-defined Networking

2.2.1 Software-defined Networking context

Software-defined Networking (SDN) has made a dramatic impact in the computer networking area. This can be attributed to its scalability, cost-effectiveness, and centralised intelligence (Javeed, Gao & Khan, 2021). SDN has demonstrated major benefits and it has been adopted by tech giant companies such as Google and Microsoft. The scalability of SDN enables organizations to easily adapt to changes within their network infrastructure in response to network needs. Since SDN is vendor-independent, an organization can configure or change the device software without having to replace the hardware (Javeed, Gao & Khan, 2021).

2.2.2 Software-defined Networking architecture

The SDN architecture consists of three planes namely, the application, control, and data plane. According to (Liu 2016), the SDN architecture "solves the bottleneck problem of large delay in response to end-to-end transmission in the big data environment by the caching function of the network node". The decoupling of the data plane from the control plane has "simplified network control and improves service deployment flexibility" (Liu 2016).

2.2.3 Software-defined Networking Security

Akhunzada et al (2015) state that "the integrity and security of SDNs remain unproven when it comes to the placement of management functionality in a single centralised virtual server." Akhunzada et al (2015) provide a way to "classify the state-of-the-art security solutions by devising a thematic taxonomy considering SDN layers/interfaces, security mechanisms, simulation environments, and security objectives."

2.3 Cyber Threat Landscape

The word cyber-security encompasses several very different considerations; thus the cyber threat landscape will also depend on a range of factors for a given country. The priorities and potentials for addressing cyber threats will vary among countries and actors within a country, as they have different national, social, economic, and organizational priorities. The cyber threat landscape is composed of eight foundational elements, some of which are technical and others non-technical, and these elements are represented as vectors of attack and response as depicted in Figure 4. This begins to answer questions concerning the nature of the cyber threat landscape faced by companies or countries as they design, develop, launch, fly and sustain intelligent space systems.

Factors affecting the cyber threat landscape of intelligent space systems include the need to secure people, processes, and tools and, as is often encountered, space systems security is not a simple issue of telecommunications security. In an era of borderless networks, one is faced with the weaponization of networking and the Internet, targeted attack on all kinds of IoT devices and commercial and Defence platforms

and the use of offensive cyber warfare or cyber effects in the Grey Zone between a declared war and intelligence operations.

At an organisational, and management level, there is often a lack of understanding of cyber threats and a lack of a security culture. Many smaller companies have little or no cyber threat understanding or protection. It has been found that many Critical Infrastructure companies lack good communication between the engineering, ICT and OT section and have poor security policies or regulations (Slay & Miller, 2007).

Figure 4: Eight vectors of attack and response (Slay & Austin, 2018)

One useful method to produce a mapping of the cyber threat landscape of a given technical context is to map the threat landscape across a range of technical and socio-technical components. This can be considered, for example, by using an old mapping of the domains of cyber security as developed by ISC^2, which are; "Access Control Systems and Methodology; Telecommunications and Network Security; Business Continuity Planning and Disaster Recovery Planning; Security Management Practices; Security Architecture and Models; Law, Investigation, and Ethics; Application and Systems Development Security; Cryptography; Computer Operations Security; and Physical Security" (ISC^2).

Thus, when considering the cyber threat landscape for an intelligent space system the obvious features of network security and telecommunications security need to be considered as do features such as the security of the applications and systems used by or drawn on by the space system and the appropriate use of cryptography. Less obvious security threats emanate from business continuity issues (how will a company respond if its satellite is hacked?) and those of cyber law and ethics.

2.4 Intelligent Security Controls for Software-defined Networking Space Systems

The Internet of Things (IoT) allows the connection of physical and virtual objects integrated with sensors and technologies that accommodate the exchange of data between devices worldwide over the Internet (Krishna et al, 2021). Smart satellite networks connect IoT devices by providing remote communication between devices and systems, as well as the extension of coverage to areas without Internet and cellular network. Smart satellite networks provide these features to IoT devices by amplifying signals that are transmitted from the sensors on the ground and transmitting them back to earth (Al-Hawawreh et al, 2021). It is estimated that IoT devices will be 25.44 billion by 2030 globally (Krishna et al, 2021). This increase in IoT devices has become a challenge for security due to unencrypted traffic, the difference in device firmware and integrated security features. These loopholes in IoT devices essentially affect technologies such as space systems which incorporate the smart or autonomous feature, leaving them vulnerable to the disruption of network services or accessibility and eavesdropping of signals transmitted between the smart satellite and the ground station to name a few issues (Shadbolt et al, 2021).

Smart satellite networks experience the challenge of inflexible network management due to their use of traditional network features and architectures (Li et al, 2016) that fail to adapt to the demands of bandwidth, computing and storage of current applications. To resolve this problem, the adoption of SDN in smart satellite networks can greatly provide scalability and flexible networking architecture that is required by critical applications such as smart satellites. SDN is one of the most promising evolving network technologies Khairi et

al, 2018) as it transforms the traditional network architecture into an independent setup where the control plane is separated from the data plane, allowing greater control and programmability functions to the network administrator (Dey & Rahman, 2019). However, these great benefits of SDN come with major security vulnerabilities. The SDN control plane poses a great vulnerability because it is the point of configuration and administration of the entire network and can be a single point of failure during an attack. There is a need to secure smart satellite networks with intelligent security controls, as they are a critical technology that supports the efficiency of IoT device connectivity.

3. Potential Solution

Based on our analysis above, intelligent security controls are thus needed for the assurance of SDN-based space systems and services. The security control proposed here is an Intrusion Detection System (IDS) with the intelligence to adapt to evolving threats. Development of the IDS is a three-stage process.

Stage 1 – Development of an SDN testbed: An SDN testbed is necessary because existing datasets such as UNSW-NB15 (Moustafa & Slay, 2015) and CSE-CIC-IDS2018 (Leevy & Khoshgoftaar, 2020) are useful for the data plane, but do not cover the control plane.

Despite the benefits of SDN mentioned in Sec. 1, SDN in its centralised form introduces a single point of failure/vulnerability. Distributed SDN can address this issue, in addition to the issues of scalability and reliability (Dixit et al, 2013). In distributed SDN, the control plane is physically distributed, but the control plane can either be logically centralised or logically distributed (Bannour et al, 2018). Over the last decade, a plethora of distributed SDN platforms have been designed and developed. For example, the Open Network Operating System (ONOS) by the Open Networking Foundation (ONF) is an open-source, extensible, modular, distributed SDN platform that implements a logically centralised control plane (Berde et al, 2014).

An SDN testbed based on ONOS will be developed, in integration with a satellite network simulator, to simulate networking scenarios. Candidate satellite network simulators include EXata, Systems Tool Kit (STK) (Li et al, 2016), Satellite Network Simulator 3 (SNS3) and OPNET (Liu et al, 2020). The attack surface of this testbed will be analysed and based on this analysis, novel attacks will be discovered.

Stage 2 – Design of an IDS for the testbed: The proposed IDS adopts the contemporary, hybrid design of combining signature-based detection and behaviour-based detection.

Against known attacks (Khairi et al, 2018; Haider et al., 2020; Malik et al, 2020; Javeed et al, 2021), features of these known attacks will be determined through feature selection and used to develop correlation rules and train supervised learning algorithms.

Against the novel attacks discovered in Stage 1, an anomaly detection scheme will be developed. The No Free Lunch Theorem (Wolpert & Macready, 1997) implies no machine learning algorithm is universally better than any other, and thus simply adopting a state-of-the-art algorithm in the anomaly detection literature is inadequate.

To support the development of the proposed algorithms, network traffic will be generated using the SDN testbed developed in Stage 1, cleaned and split into training, validation and test datasets. Deep neural network architectures will be standard as deep features are efficient to extract and as effective as handcrafted features. Furthermore, continual learning of evolving threats will be achieved through adaptive model updates (Stocco & Tonella, 2020). To cope with the large data volume and processing power requirements, National Computational Infrastructure's Gadi supercomputer will be used.

Stage 3 – Testing of the IDS: k-fold cross-validation (Goodfellow et al, 2016) will be used to characterise the performance of the algorithms developed in Stage 2 in terms of the confusion matrix (for classification of attack types), detection accuracy, detection precision (capturing the false positive rate), detection recall (capturing the false negative rate), area under the receiver operating characteristic curve, computational complexity as well as memory consumption. Table 1 defines the metrics that will be used for testing the IDS.

Table 1: Evaluation metrics used to assess the performance of the proposed DL-based IDS

Metrics	Description
Confusion matrix	This measures the performance of a model based on TP, FP, TN and FN values.
Accuracy (A)	Returns the probability that an instance is correctly classified by the IDS.

Precision (PR)	The ratio of correctly classified attack instances (TP), in front of all classified flows (TP+FP).
Recall (RC)	The ratio of correctly classified attack instances (TP), in front of all generated flows (TP+FN).
Loss	This is an indication of the model's prediction ability. If the model has an excellent prediction, the loss value is zero or closer to zero; else, the loss value is far greater than zero.
True Positive Rate (TPR)	Measures how often the model correctly predicts positive instances.
False Positive Rate (FPR)	Measures how often the model incorrectly predicts positive instances.
ROC AUC curve	Measures the false alarm rates against the actual predicted rates, showing how well a model can make a distinction among classes.
Central Processing Unit (CPU) usage (%)	Measures the capacity of the CPU used in training the model.
Train and Test Time	This is the time taken for training and testing the model.

4. Conclusion

Space systems are important to critical systems such as the military, weather services, and global timing. Remote communication between devices and systems, as well as the extension of coverage to areas without Internet and cellular network, are enhanced by intelligent space systems. Due to space systems using terrestrial networks, there are limitations such as inflexibility and complex network management. Thus, we propose the integration of SDN and space systems to take advantage of the advancements provided by SDN. SDN is a communication network approach that enhances a network's flexibility, scalability, and programmability. However, SDN and space systems, individually and when combined, are at risk of cyber-attacks that intend to maliciously compromise the confidentiality, integrity, and availability of SDN and space systems. Thus, we propose intelligent security controls that will protect SDN-based space systems from the rising novel cyber-attacks.

References

Aguirre, M.A. (2013) 'Introduction to Space Systems: Design and Synthesis', *Springer*, 5(3), pp. 248–253. Available at: https://doi.org/10.1007/978-1-4614-3758-1.

Akhunzada, A. *et al.* (2015) 'Securing software defined networks: Taxonomy, requirements, and open issues', *IEEE Communications Magazine*, 53(4), pp. 36–44. Available at: https://doi.org/10.1109/MCOM.2015.7081073.

Al-Hawawreh, M., Moustafa, N. and Slay, J. (2021) 'A threat intelligence framework for protecting smart satellite-based healthcare networks', *Neural Computing and Applications*, 5. Available at: https://doi.org/10.1007/s00521-021-06441-5.

Bannour, F., Souihi, S. and Mellouk, A. (2018) 'Distributed SDN Control: Survey, Taxonomy, and Challenges', *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 333–354. DOI: 10.1109/COMST.2017.2782482.

Berde, P. et al. (2014) 'ONOS: Towards an Open, Distributed SDN OS', *Proceedings of the Third Workshop on Hot Topics in Software Defined Networking, HotSDN* '14, New York, NY, USA, Association for Computing Machinery, pp. 1–6. DOI: 10.1145/2620728.2620744.

Bertaux, L. *et al*. (2015) 'Software defined networking and virtualization for broadband satellite networks', *IEEE Communications Magazine*, 53(3), pp. 54–60. Available at: https://doi.org/10.1109/MCOM.2015.7060482.

Bradbury, M. *et al.* (2020) 'Identifying Attack Surfaces in the Evolving Space Industry Using Reference Architectures', *2020 IEEE Aerospace Conference*, pp. 1–20. Available at: https://doi.org/10.1109/AERO47225.2020.9172785.

Dey, S.K. and Rahman, M.M. (2019) 'Flow Based Anomaly Detection in Software Defined Networking: A Deep Learning Approach With Feature Selection Method', *4th International Conference on Electrical Engineering and Information and Communication Technology, ICEEICT*, pp. 630–635. Available at: https://doi.org/10.1109/CEEICT.2018.8628069.

Dixit, A. *et al.* (2013) 'Towards an Elastic Distributed SDN Controller', in *Proceedings of the Second ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking*, 43(4), pp. 7–12. Available at: https://doi.org/10.1145/2491185.2491193.

Falco, G. (2018) 'Cybersecurity Principles for Space Systems', *Journal of Aerospace Information Systems*, (December 2018). Available at: https://doi.org/10.2514/1.I010693.

Ferrús, R. *et al.* (2016) 'SDN/NFV-enabled satellite communications networks: Opportunities, scenarios and challenges', *Physical Communication*, 18, pp. 95–112. Available at: https://doi.org/10.1016/j.phycom.2015.10.007.

Goodfellow, I., Bengio, Y., and Courville, A. (2016) 'Deep Learning', *MIT Press.* Available at: http://www.deeplearningbook.org.

Khairi, M. et al. (2018) 'A Review of Anomaly Detection Techniques and Distributed Denial of Service (DDoS) on Software Defined Network (SDN)', *Engineering, Technology & Applied Science Research*, 8. Available at: https://doi.org/10.48084/etasr.1840.

Haider, S. *et al.* (2020) 'A Deep CNN Ensemble Framework for Efficient DDoS Attack Detection in Software Defined Networks', *IEEE Access*, 8, pp. 53972–53983. Available at: https://doi.org/10.1109/ACCESS.2020.2976908.

Javeed, D., Gao, T. and Khan, M.T. (2021) 'SDN-Enabled Hybrid DL-Driven Framework for the Detection of Emerging Cyber Threats in IoT', *Electronics*, 10(8), pp. 1–16. Available at: https://doi.org/10.3390/electronics10080918.

Kaloudi, N. and Li, J. (2020) 'The AI-Based Cyber Threat Landscape : A Survey', *ACM Computing Surveys*, 53(1), pp. 1-34. Available at: https://doi.org/10.1145/3372823.

Krishna, R.R. *et al.* (2021) 'State-of-the-Art Review on IoT Threats and Attacks: Taxonomy, Challenges and Solutions', *Sustainability*, 13(16), pp. 1–45. Available at: https://doi.org/10.3390/su13169463.

Leevy, J.L. and Khoshgoftaar, T.M. (2020) 'A survey and analysis of intrusion detection models based on CSE-CIC-IDS2018 Big Data', *Journal of Big Data*, 7(1), p. 104. Available at: https://doi.org/10.1186/s40537-020-00382-x.

Li, T. *et al.* (2016) 'Using SDN and NFV to Implement Satellite Communication Networks', *Proceedings - 2016 International Conference on Networking and Network Applications, NaNA 2016*, pp. 131–134. Available at: https://doi.org/10.1109/NaNA.2016.22.

Liu, M. *et al.* (2020a) 'Large-Scale Small Satellite Network Simulator: Design and Evaluation', *2020 3rd International Conference on Hot Information-Centric Networking, HotICN 2020*, pp. 194–199. Available at: https://doi.org/10.1109/HotICN50779.2020.9350838.

Liu, Y. *et al.* (2020b) 'A Shared Satellite Ground Station Using User-Oriented Virtualization Technology', *IEEE Access*, 8, pp. 63923–63934. Available at: https://doi.org/10.1109/ACCESS.2020.2984485.

Liu, Z. *et al.* (2018) 'Satellite Network Architecture Design Based on SDN and ICN Technology', *2018 8th Proceedings of 2018 IEEE 8th International Conference on Electronics Information and Emergency Communication, ICEIEC*, pp. 124–131. Available at: https://doi.org/10.1109/ICEIEC.2018.8473548.

Malik, J. *et al.* (2020) 'Hybrid Deep Learning: An Efficient Reconnaissance and Surveillance Detection Mechanism in SDN', *IEEE Access*, 8, pp. 134695–134706. Available at: https://doi.org/10.1109/access.2020.3009849.

Manulis, M. et al. (2020) 'Cyber security in New Space: Analysis of threats, key enabling technologies and challenges', International Journal of Information Security, 20(3), pp. 287–311. Available at: https://doi.org/10.1007/s10207-020-00503-w.

Moustafa, N. and Slay, J. (2015) 'UNSW-NB15: A Comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)', *2015 Military Communications and Information Systems Conference, MilCIS*, pp. 1–6. Available at: https://doi.org/10.1109/MilCIS.2015.7348942.

Ormrod, D., Slay, J. and Ormrod, A. (2021) 'Cyber-Worthiness and Cyber-Resilience to Secure Low Earth Orbit Satellites', *International Conference on Cyber Warfare and Security*, pp. 257-266. Available at: https://www.proquest.com/docview/2505729841?pq-origsite=primo.

Pavur, J. and Martinovic, I. (2020) 'SOK : Building a Launchpad for Impactful Satellite Cyber-Security Research', pp. 1–32. Available at: https://doi.org/10.48550/arxiv.2010.10872.

Plotnek, J. and Slay, J. (2022) 'Space Systems Security: A Definition and Knowledge Domain for the Contemporary Context', *Journal of Information Warfare*, 21(3), pp. 103–119. Available at: https://www.jinfowar.com/journal-issue/volume-21-issue-3.

Pradhan, A. and Mathew, R. (2020) 'Solutions to Vulnerabilities and Threats in Software Defined Networking (SDN)', *Procedia Computer Science*, 171, pp. 2581–2589. Available at: https://doi.org/10.1016/j.procs.2020.04.280.

Shadbolt, L. (2021) *Technical Study Satellite Cyberattacks and Security*, *HDI Global Specialty SE*. Available at: https://www.hdi.global/infocenter/insights/specialty/technical-study/.

Slay, J. and Austin, G. (2018) 'Development in Training and Education for Australian Cyber Security : Filling the Gaps', *Journal of The Colloquium for Information System Security Education (CISSE)*, 5(2), pp. 1–27. Available at: https://cisse.info/journal/index.php/cisse/article/view/80.

Slay, J. and Miller, M. (2007) 'Lessons Learned from the Maroochy Water Breach', *Critical Infrastructure Protection*, 253, pp. 73–82. Available at: https://doi.org/10.1007/978-0-387-75462-8_6.

Stocco, A. and Tonella, P. (2020) 'Towards Anomaly Detectors that Learn Continuously', *ISSREW 2020*, pp. 201-208. Available at: https://doi.org/10.1109/ISSREW51248.2020.00073.

The International Information Systems Security Certification Consortium, Inc., (ISC)², Certification Programs, CISSP Domains. Available at: https://www.isc2.org/cissp-domains/default.aspx.

Wolpert, D. H. and Macready, W. G. (1997) 'No free lunch theorems for optimization', *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82. DOI: 10.1109/4235.585893.

Masters Research Papers

How Does the Tallinn Manual 2.0 Shed Light on the Threat of Cyber Attacks against Taiwan?

Chih-Hsiang Chang
Graduate, Institute of Interdisciplinary Legal Studies, National Taiwan University

r07a41028@ntu.edu.tw

Abstract: This paper will identify possible unsettled issues when applying jus ad bellum and jus in bello to case scenarios based on China's cyber operations against Taiwan, pursuant to the rules of international law governing cyber or military operations attributable to States reflected in Tallinn Manual 2.0. This paper will argue that because of Taiwan's legal international status as a sovereign State, the different responsive actions it may take, should it be faced with any such aggressive cyber or military attack, may be considered controversial. This paper will then identify the possible legal issues that may pertain under current international law, should any such armed conflict occur between China and Taiwan.

Keywords: cyber attack, cyber armed conflict, international humanitarian law, the Tallinn Manual 2.0, Taiwan Strait Crisis, right of self-defence.

1. Introduction

As a regional power in Asia, China does not shy away from its ambition to "reclaim Taiwan" as its territory, and recent military aggression has inevitably pushed Taiwan and China to the edge of another Taiwan Strait Crisis. Apart from the constant threat of kinetic attacks, Taiwan has also become a hotspot for frequent disinformation and cyber operations in both the private and public sectors, which have often been identified as suspicious activities, conducted by hackers from China.

Russia's invasion of Ukraine in 2022 is a typical case of how multiple cyber operations, both prior to and during an assault, can be deployed either to hike up geo-political tensions, or as a precursor to actual armed conflict. Consequently, should China adopt similar tactics towards Taiwan, it may use similar tactics and initiate cyber attacks against Taiwan, whether during peacetime or in an actual war. Thus, it is important to examine what possible legal challenges Taiwan would be faced with, given its current international status, while dealing with the threats from cyberspace, particularly since it is often treated by other sovereign States as merely a de facto self-governing entity and not a fully-recognised State.

This paper will answer the following questions:

1. How will Taiwan be governed by the Tallinn Manual 2.0 Rules regarding armed conflict?
2. What legal issues may arise if Taiwan applied the Tallinn Manual 2.0, should it be targeted by Chinese cyber attacks?
3. Combating the cyber attacks, what are possible lawful responses for Taiwan to choose from under the framework of international law?

This paper's methodology will be as follows: (i) a brief review of the controversy over Taiwan's current international legal status, (ii) a contemporary history and geo-political power struggle assessment will be described, (iii) the controversies in terms of applying IHL to Taiwan in terms of armed conflict will be explained, (iv) publicly-known cyber operations against Taiwan will be examined, (v) case scenarios, based on the rules of the Tallinn Manual 2.0, will be described in order to determine how malicious cyber operations shall be governed with regard to international law, and (vi) critical legal issues regarding Taiwan's exercise of its rights to self-defence in the event of a cyber attack will be discussed.

2. The Legal Status of the Republic of China (Taiwan)

To discuss how the international humanitarian law shall be applied if Taiwan becomes a party of an armed conflict, the first priority would be to determine whether Taiwan may be deemed to be a sovereign State.
Taiwan and the Republic of China (ROC) are practically — and diplomatically — synonymous in the eyes of today's international society; however, there is a divergence between its de jure and de facto legal status which cannot be ignored.

2.1 Historical background

After the implementation of the Treaty of Shimonoseki (1895) between the Empire of Japan and Qing China,

Taiwan (the Formosa Island and the Pescadores) fell under the rule of Japan as an oversea colony until Japan surrendered to the Allied Powers in 1945, when it renounced its rights over Formosa and Pescadores in the San Francisco Peace Treaty of 1951 (Chen, 1998).

Following General Order No.1 by General Douglas MacArthur, Taiwan experienced a post-World War II military occupation by ROC, whose government (the government formed by the Chinese Nationalist Party, which is also known as KMT) was led by Generalissimo Chiang Kai-Shek. However, following its defeat in a civil war (1945-1949) by the Chinese Communists Party, Chiang fled to Taiwan with his army and imposed perpetual martial law on the island for 38 years in the name of the Nationalist Chinese authorities (1949 to 1987), which caused the islanders to suffer what they called "the white terror" employed by an authoritarian regime, notorious for violating human rights (Chen, 1998).

However, as a consequence of the United Nations General Assembly (UNGA) Resolution, 2758 of 1971, which not only terminated the membership of the Republic of China (ROC), but also resulted in most members of the international community recognizing the People's Republic of China (PRC) as the only representative of China instead of ROC. Thus, during the 1970s and 1980s, faced with the dual challenge to its Chinese representational legitimacy and relentless local campaigns in Taiwan for civil rights, the KMT leadership eventually conceded to a transformation to democracy through "Taiwanization" (Chen, 1998). During the 1990s, new articles of the Constitution, allowing for general elections of legislators and the president of the central government, were instituted. In 1996, the Taiwanese were able to enjoy full political democratic power. Then in 2000, came the remarkable peaceful transfer from the KMT to the Democratic Progressive Party (DPP) (Chiang & Hwang, 2008).

2.2 Controversies surrounding Taiwan's legal status

Theories regarding Taiwan's legal status fall into three categories.

1. That Taiwan should be part of China through "acquisitive prescription"; this has long been advocated by the PRC through its "one China principle". Hence, PRC treats Taiwan as a "renegade province" through State practice.
2. That ROC (Taiwan) "has emerged as an independent State" (Chiang & Hwang, 2008); hence it should be considered a sovereign State, even though it is officially called the Republic of China. Although it has never officially declared its independence - a requisite under the "Declaratory Theory of Recognition" – its supporters believe that its statehood is self-evident through State practice.
3. That Taiwan is an international condominium of the Allied Powers; therefore, the issue regarding its sovereignty is not an internal affair of China and should be settled through peaceful multilateral actions. They reject the PRC's claim, and deny the legitimacy of the ROC since Generalissimo Chiang was only acting as a delegate on behalf of those States that were parties to the San Francisco Treaty; it is those States that should collectively either decide the future of Taiwan, or transfer it to the United Nations Trusteeship Council.

Apart from theoretical disputes, the official position of the ROC has changed dramatically. Before 1973, it was much closer to that of PRC than to other major powers. The main difference was that ROC claimed to be the only legitimate government of China. However, after PRC obtained the representation of China in the UN in 1973, ROC limited its sovereign claim to those territories that had been under its effective control since 1949 – Taiwan (Formosa), Penghu (the Pescadores), Kinmen (Quemoy) and Matsu (Chiang & Hwang, 2008); this remains its official position to date.

Prior to 1973, the international legal mainstream and major powers favoured the third position; however, substantial changes have reshaped the latter's foreign policies since then; also, following Taiwan's democratization, many scholars agree that Taiwan is "a self-governing entity whose legal status remains indeterminate", thus it is prevented from possessing the usual rights, obligations, and immunities of a recognized independent State, even though it is both "edging towards formal separation" and is "functioning as a de facto State" (Crawford, 2006).

2.3 Tallinn Manual 2.0 and Taiwan: The Application of the Law in Terms of a Cyber Operation in nexus with an Armed Conflict

To identify whether a cyber operation in nexus with an armed conflict could occur between Taiwan and China,

and to examine the legal issues regarding international humanitarian law should it do so, it is crucial to define what an armed conflict is and what its nature might be.

Pursuant to Common Article 2 of the Geneva Conventions of 1949 and the accompanying International Committee of the Red Cross (ICRC) Commentary, an international armed conflict (IAC) exists whenever there is resort to hostile armed force between two States. The threshold of a non-international armed conflict (NIAC) is much higher. Common Article 3 covers a "conflict not of an international character occurring in the territory of one of the High Contracting Parties"; however, it lacks a clear definition. Nevertheless, the International Criminal Tribunal for the Former Yugoslavia (ICTY) provided a standard in the Tadić judgement, which required two conditions: a protracted armed violence (i) between government authorities and organized armed groups, or (ii) between such groups within a State.

Regarding cyber operations, the Tallinn Manual 2.0 states that "cyber operations executed in the context of an armed conflict are subject to the law of armed conflict" (Rule 80); however, two requisites must be satisfied: (i) the existence of cyber operations, and (ii) a nexus between the cyber activity in question and the conflict for the law of armed conflict to apply to that activity. Regarding the classification, the definitions of a cyber IAC and a NIAC are included in Rules 82 and 83, and they must fully reflect the principles of Common Articles to the Geneva Conventions of 1949 and the standard from the Tadić judgement.

2.4 If Taiwan becomes involved in an armed conflict with China, would it be considered to be international or non-international?

The debate about whether Taiwan's statehood plays a significant role in terms of International Humanitarian Law (IHL). If a position (regardless of reasons) is adopted by arguing that Taiwan is not an independent State, then the requirement of "a protracted armed violence", implying greater extent, duration, or intensity of hostilities, must be satisfied. This approach would not favour Taiwan if it were to face imminent hostilities that endanger the safety of its people, especially when one of China's strategies is to initiate a surprise attack to consolidate its military advantage and create a fait accompli, which may leave other stakeholders, or the UN, to find it too hard, or too late, to intervene.

In the cyber context, a similar problem regarding the threshold of NIAC also exists. In the Tallinn Manual 2.0, the Experts noted that "the precise parameters of the phrase 'in the context of' are less clear in a non-International armed conflict," because "a State retains certain law enforcement obligations and rights with respect to its territory, notwithstanding the armed conflict" (Commentary to Rule 80). Considering there are more uncertainties about whether certain cyber operations qualify as cyber attacks, thereby meeting the standard of armed attack under Article 51 of the UN Charter (see Sections 3.2 & 3.3), treating Taiwan as a non-sovereign State increases the difficulties of applying IHL on cyber attacks. Therefore, if applying the Declaratory Theory stricto sensu, IHL may not favour Taiwan should it face imminent hostilities that endanger the security and welfare of its people.

2.5 Cyber attacks Governed by International Humanitarian Law

i. The definition of the term "cyber attack"

Rule 92 of the Tallinn Manual 2.0 defines a cyber attack as "a cyber operation, whether offensive or defensive, that is reasonably expected to cause injury or death to persons or damage or destruction to objects", which follows Article 49(1) to Additional Protocol I of the 1949 Geneva Conventions (AP I), stating "acts of violence against the adversary, whether in offense or defence."

The Tallinn Manual 2.0 also states that "any cyber operation rising to the level of an armed attack in terms of scale and effects pursuant to Rule 71 qualifies as a 'use of force' if it is conducted by or otherwise attributable to a State (Commentary to Rule 69). In addition, the Experts "unanimously concluded that some cyber operations may be sufficiently grave to warrant classifying them as an 'armed attack' within the meaning of the Charter." (Commentary to Rule 71). However, they agreed that de minimis damage or destruction does not meet the threshold of harm required by Rule 92 (Commentary to Rule 92). In summary, the Tallinn Manual 2.0 does not deny that certain cyber operations alone can be armed attacks and therefore constitute armed conflict. In compliance with IHL, the Tallinn Manual 2.0 proposes numerous Rules to limit the scope of lawful military objectives that cyber attacks may target.

Two of its important principles are discussed through a case study, (i) civilian infrastructure is protected by existing IHL rules, particularly the principles of distinction, proportionality and precautions in attack, and (ii)

civilian data shall be protected from significant disruption. Intentional violations of these, directed against civilian objects may be war crimes (Article 8(2)(b)(ii) of the Rome Statute (1998).

ii. The unsettled issues in the Tallinn Manual 2.0 relevant to targets attacked by cyber means

Before applying the Tallinn Manual 2.0 on scenarios in Taiwan, note that several issues remain unsettled due to divisions among State's positions and practices concerning the definition of "damage or destruction to objects" in Rule 92, based on Article 52(2) of AP I.

The first concerns whether "interference by cyber means with the functionality of an object" constitutes "damage or destruction to objects" (Rule 92). The Experts were broadly split on this issue. The majority insisted that only when restoration of functionality requires replacement of physical components, does interference qualify as damage. The second position, held by some members of the majority, was that such interference "extends to situations in which reinstallation of the operating system or of particular data is required in order for the targeted cyber infrastructure to perform the function for which it was designed". The third position, adopted by a few Experts, held that how an object is disabled is irrelevant; instead, the loss of usability of cyber infrastructure constitutes damage that qualifies as a cyber attack if cyber infrastructure in question is targeted (Commentary to Rule 92 of Tallinn Manual 2.0).

The second unsettled issue is whether data is an "object" under Rule 92, on which the Experts widely diverged. In the view of the majority of the Experts, data is intangible and therefore "neither falls within the ordinary meaning of the term object, nor comports with the explanation of it offered in the ICRC Additional Protocols 1987 Commentary". On the contrary, a minority of the Experts was of the opinion that, for the purposes of targeting, certain data should be regarded as an object. Also, for the minority, it is based on the underlying object and purpose of Article 52 of Additional Protocol I, instead of the nature of harm, the key factor should be "the severity of the operation's consequences". (Commentary to Rule 92)

2.6 Case study: a basic legal analysis of a possible cyber attack against Taiwan

In Section 3.3, principles and rules from the Tallinn Manual 2.0 will be applied to different scenarios, if cyber operations are launched against Taiwan. This case study will proceed based on the assumption that Taiwan is an independent sovereign State. Two questions are posed, (i) if there is a nexus between the cyber operation and the existing armed conflict, can the cyber operation constitute a cyber attack as defined by Rule 92, depending on its scale and effect (Rule 69)? And (ii) is the purported cyber attack against an unlawful target under IHL, according to Tallinn Manual 2.0?

iii. Submarine cables near Taiwan: a possible trigger to armed conflict

In February 2023, submarine cables providing the Internet service in Matsu were damaged, causing a temporary shut-down of the Internet connection. This was not the only incident in recent years, and most have occurred due to civilian fishing vessels, which, according to media reports, have been Chinese dredgers.

It is generally agreed that, should China launch military operations to change the status quo by damaging submarine cables, Internet connectivity on and to Taiwan will be severely restricted, or cut entirely. This could encourage China to create a 'war-mist' during a beginning phase of an invasion. One vulnerable site is in the Luzon Strait near the Philippines, where all cables connecting Hong Kong, Taiwan, South Korea and Japan, are situated, which could disrupt service in the whole of northeast Asia, thus potentially constituting a cyber-attack under Rule 92, violating Article 113 of the United Nations Convention on the Law of the Sea (UNCLOS). Therefore, if China disrupts submarine cables providing most of the Internet service for Taiwan by cyber means, assuming it is technically feasible, do the operations constitute a cyber attack under Rule 92?

Clearly, such operations are violation of States' responsibilities for cable protection if the targeted submarine cables located on the high seas (Article 113 of UNCLOS). The Tallinn Manual 2.0 Experts agreed, in that "the infliction of damage to cables by a State is prohibited as a matter of customary international law since doing so would run contrary to the object and purpose of the law governing submarine cables". Furthermore, they agreed, "there is no legal basis to cut another State's submarine fibre optic cable in order to reduce trans-continental Internet traffic in times of tension" (Commentary to Rule 54).

However, a crucial point is whether the cyber operation caused "damage or destruction to objects" pursuant to Rule 92. If the positions supported by the majority of the Experts were adopted, unless the operations required the replacement of physical components of optical fibres, such as materials for the core, the cladding or the coating, the interference itself would not qualify as damage. However, if the minority's position were followed,

the consequence of successfully blocking the Internet service of Taiwan and other neighbouring countries, resulting in significant economic loss, could be considered a hostile cyber attack.

Even if such a cyber operation did constitute a cyber attack, because communications for military use often share the same cables with commercial communications, it is hard to distinguish the one from the other. Thus the principle of "dual-use objects" renders submarine cables as lawful targets (Rule 95, Tallinn Manual 2.0) while a traditional interpretation of IHL sets a lower bar regarding the principles of distinction (Rule 93, the Tallinn Manual 2.0) and the proportionality (Rule 113) in the case of targeting submarine cables. This aligns with historical precedents and States' practices regarding the application of law of armed conflict on kinetic attacks against submarine cables (Harbin, 2021).

ii. Industrial Control System (ICS) of private companies: the vulnerability of the semiconductor-manufacturing industry

So far, suspicious Chinese cyber operations that have been detected were denial-of-service attacks (DDoS) against government websites, or the dissemination of rumours or disinformation, which mostly "merely cause[d] inconvenience or irritation to the civilian population" and have created minor temporary disruptions, which fall short of armed attacks (Commentary to Rule 92, the Tallinn Manual 2.0) Thus, their scale and effects were largely immaterial (Rule 69, the Tallinn Manual 2.0). However, it has been deemed advisable for private companies to evaluate the potential cybersecurity risks, especially regarding the nexus between the semiconductor industry and national security.

The Taiwan Semiconductor Manufacturing Company (TSMC) reportedly controls more than 60% of the world's semiconductors and over 90% of the most advanced ones; it alone contributes 15% of Taiwan's GDP. Therefore, from China's perspective, to sabotage it would not only undermine Taiwan's economic strength, but also cause negative influence on other States, such as the US. Furthermore, Industrial Control System (ICS) is a collective term used to describe control systems and associated instrumentation, including the devices, systems, networks, and controls used to operate and automate industrial processes. This mostly include Supervisory Control and Data Acquisition (SCADA) systems and Distributed Control Systems (DCS), most of which are highly-automated and rely heavily on ICSs, particularly in the semiconductor manufacturing industry. For example, a statistical process control is crucial for semiconductor manufacturers to monitor process variability and detect outliers.

In order to gain leverage, or to coerce Taiwan to submit to its political agenda, if China launched cyber operations against the SCADA systems of TSMC, or the main Taiwanese semiconductor manufacturers, by deleting, or altering technical data, causing significant economic loss and a stock market crash; whether this would satisfy the definition of cyber-attack under Rule 92 is hard to establish due to the lack of consensus about such data falling under that rule; however, it is worth referring to a broader interpretation of Article 52 of AP I, which some scholars agree that "data is an object under IHL, therefore it may be a military objective" (Mačák, 2015).

The Vienna Convention on the Law of Treaties (1969) does not apply to norms of customary law. Therefore, it is possible that a norm exists in parallel in both treaty law and customary law, which is clarified in the Nicaragua judgement of the International Court of Justice (ICJ). Therefore, Article 52(2) can be "described as a 'fundamentally norm-creating' treaty rule of the kind the ICJ considered to be eligible of evolving into custom in the North Sea Continental Shelf cases" (Mačák, 2015). Therefore, by adopting the minority's position, should a destructive cyber attack be initiated against crucial civilian data by cyber means, there is a possibility that, depending on the scale and effects manifested (Rule 69, the Tallinn Manual 2.0), cyber operations against TSMC, or the whole Taiwanese semiconductor industry, may qualify as a cyber attack.

If this conclusion is accepted, the next step would be to examine whether TSMC and other similar companies are lawful military objectives. Although the semiconductor manufacturing industry is a part of the private sector, it is crucial to examine whether this satisfies the "dual-use object" standard in Rule 95. In contemporary warfare, semiconductors are of vital strategic significance, because no military equipment and weaponry, such as satellites, drones, guided missiles and jets, can function without sophisticated electronic chips.

In the Commentary to Rule 100 of the Tallinn Manual 2.0, two scenarios were given. The first one is a factory producing computer hardware or software under contract to the enemy's armed forces; the second is the enemy State's oil export industry, on which its war effort depends heavily. All the Experts agreed on the computer factory case, which by its use is a military objective; however, in the second example, the majority of the Experts rejected the concept that "war-sustaining objects" are lawful military objects, as stated in the US Department of Defence Law of War Manual.

Despite this divergence of opinion, considering the indispensable role electronic chip-producing plays in modern weapon systems, semiconductor companies, such as TSMC, may qualify both as the producer for its armed forces, and as war-sustaining objects, which would render their chip manufacturers both lawful, and vulnerable, military objectives during a war.

3. A dilemma: legality vs legitimacy

Even though the theoretical disagreements surrounding the sovereignty controversies of Taiwan do not substantially affect its active role as an influential political entity in the real-world of international politics, they would make a difference if it chooses to claim the right of self-defence, or to adopt countermeasures to protect its cyber-security as an independent State under international law. Such issues cannot afford to be overlooked, especially in such times of heightened tension.

iv. Self-defence and customary international law

The doctrine of a State's right to self-defence in customary international law was crystalized after the Caroline incident, and the rules in terms of exercising this right were stipulated in Article 51 of the UN Charter. While Article 2(4) strictly prohibits the "threat or use of force in international relations", Article 51 covers one of the rare exceptions for a State to legally employ the use of force, which is explicitly invoked where a UN Member State has suffered an "armed attack" — i.e. a grave form violation of Article 2(4)'s prohibition on the use of force, which was explicitly distinguished from the term 'use of force' in the ICJ Nicaragua judgment.

v. Self-defence in the cyber context: the issue of whether a cyber operation constitutes an armed attack

It was unanimously agreed by the Tallinn Manual 2.0 Experts that some cyber operations "may be sufficiently grave to warrant classifying them as an armed attack" within the meaning of the UN Charter. Consequently, the Experts concurred that the right to employ force in self-defence extends to those armed attacks that are conducted solely through cyber operations, which do not necessarily involve the employment of weapons (Rule 103) (Commentary to Rule 71, the Tallinn Manual 2.0).

In the Tallinn Manual 2.0, the threshold required for cyber operations constituting armed attack is high. For instance, the Experts expressed a "general agreement that cyber operations that merely cause inconvenience or irritation to the civilian population do not rise to the level of attack" (Commentary to Rule 92); additionally, the following activities in cyberspace do not qualify as armed attacks: "acts of cyber intelligence, gathering and cyber theft, as well as cyber operations that involve brief or periodic interruption of non-essential cyber services" (Commentary to Rule 71).

3.1 Taiwan to exercise the right of self-defence against cyber attacks by China

vi. Self-defence as a "limited-to-States" right

If Taiwan is not recognised as an independent sovereignty, it complicates the legal relationships between stakeholders, thus rendering the existing sovereignty issues surrounding Taiwan to be more challenging, particularly when pursuing the sensu stricto legality under international law.

Firstly, ROC (Taiwan) cannot claim the right of self-defence because this is limited to sovereign States. Although, in order to protect its people from imminent danger of threat, or use of force, its authorities may argue that the international community should follow the doctrine of "Responsibility to Protect" and take immediate action, military or not, to justify humanitarian intervention from other States by referring to the precedent case of Kosovo; however, this legal doctrine is still highly contested in international law. Additionally, it is generally observed that, unless abhorrent atrocities, such as grave war crimes or genocide are committed repeatedly, most States will tend not to actively support the claims of groups seeking self-determination, and adopt the principle of non-intervention, because they are regarded as the internal affairs of a State.

Secondly, the inconvenient truth of the awkward legal position of the ROC armed forces would be brought to the table. The military forces as a part of the ROC establishment could be regarded as an armed group, which does not act on behalf of a sovereign State, because ROC is no longer regarded as the legitimate regime representing China by the majority of the international community. Although the Taiwanese authorities may claim that as a party of the Treaty of Peace Between the Republic of China and Japan (also known as the Treaty of Taipei, 1952) and a member of the Allied Powers, the ROC and other parties to the San Francisco Peace Treaty, bear an obligation to maintain international peace and security in respect to Article 2 of the UN Charter;

however, this argument does not provide the legal basis to exercise the right of collective self-defence by other States, such as Japan or the US, since Taiwan is a “non-State actor” (Himes & Kim, 2021).

Consequently, in international law, as long as Taiwan is regarded merely as a self-governing entity, a roundabout way to deal with the legalities regarding the countermeasures it may undertake must be proposed. Considering the UN has no systematic integrated policy for engaging, or intervening, with de facto authorities, but only separate guidelines, based on a case-by-case approach, the ambiguity of Taiwan’s status can only bring further legal uncertainty, which could result in the interests of its people being neglected. An alternative solution might be a "quasi-declaration of independence" (Chiang & Hwang, 2008) by which Taiwan implies its intention to be regarded as a sovereign State, which should settle most of the abovementioned legal issues and result in the best outcome.

vii. Taiwan’s strike back: by armed force or by civilians?

Countermeasures may be adopted by a State when two perquisites are satisfied: (i) that the breach of an international obligation owed to the injured State is established, and (ii) that the purported wrongful act is attributable to the responsible State. Launching an armed attack against another State is certainly a violation of the rule prohibiting the use of force under customary international law. However, in the cyber context, if a State decides to "hack back" as a lawful countermeasure against cyber attacks, it is required to identify a violation of international law by the hacker group in order to determine whether, and how, the group's activities were affiliated with the responsible State (Schmitt, 2013).

In Taiwan’s scenario, a cyber countermeasure could be conducted by digital ‘armed forces’. In fact, the Information Communication Electronic Force Command (ICEF), which was inaugurated on July 1, 2017, is crucial to Taiwan’s offensive cyber operations. Its purpose is to consolidate communication, cyber, and electronic warfare units under one authority. Nevertheless, observing the current policy adopted by President Tsai’s administration, it is more likely that the ICEF will develop defensive strategies and tactics, rather than to proactively hack back.

On the other hand, a group of “volunteer hackers” may become involved in the cyber armed conflict, since such participation is not unlawful under IHL. For instance, after the first day of Russia’s invasion, Ukrainian Vice Prime Minister Fedorov publicly called on volunteer hackers and announced a list of thirty-one websites of Russian banks, corporations and government agencies as being valuable targets. Thus, they reportedly amassed an “IT army” of more than 400,000 volunteers. According to some Chinese official reports, China has detected a hacker group called “Green Spot”, which allegedly targets its own State organs, including military institutions, embassies, financial institutions, and its nuclear industry.

However, according to the Tallinn Manual 2.0, the Experts remained divergent regarding the status of patriotic hackers. While the majority agreed that civilians retain civilian status, even if they directly participate in cyber hostilities; however, they themselves may be lawfully targeted. Therefore, unless they qualify as participants en masse in a levée, they will lack combatant immunity for their actions. Furthermore, a minority of the Experts rejected the notion that such individuals would benefit from the protection of Geneva Conventions III or IV, because they qualify neither as combatants or civilians (Commentary to Rule 91).

4. Conclusion

This paper describes the legal risks and uncertainties that are inherent in Taiwan’s uneasy relationship with China, which seems to be perpetually poised to spill over into either cyber, or armed, conflict, thereby threatening its economic stability and the well-being of its people. Because of its peculiar historical circumstances, and given the delicate balance of world politics, Taiwan must take a nuanced approach in order to establish a stable global position. However, the inconsistency between the de jure legal status and de facto international politics, is immense, which means that it would be difficult for the international community to respond should such an unwanted situation arise. At the same time, Taiwan must continue to strive to find a way to protect its people, for the longer it lingers in a state of uncertainty the more risk its people will have to face.

Should such a crisis arise with China, legal barriers could prove to be highly intractable obstacles. They could also present as unstable factors, both for Taiwan itself, and for other States that will have to determine whether, or not, to intervene, since the lawfulness of a State’s actions will play a key role in adopting any response to an armed conflict.

This paper suggests it is likely that Taiwan will continue to be trapped in an obscure zone within the realm of international law. Consequently, it calls for an overall scrutiny and thorough examination of its position in the international community in order to determine how it might develop an integrated legal strategy in response to a possible cyber, or armed, conflict across the Taiwan Strait, irrespective of whether it be kinetic or cyber in nature.

References

Chen, L. C. (1998). "Taiwan's Current International Legal Status", 32 NEW ENG. *L. REV, 675.*

Chiang, H. C., & Hwang, J. Y. (2008). "On the statehood of Taiwan: A legal reappraisal", *The "one China" dilemma*, pp 57-80.

Crawford, J. (2006). *The creation of states in international law*, Oxford University Press.

Harbin, D. (2021). "Targeting submarine cables: new approaches to the law of armed conflict in modern warfare", *Military Law Review, 229(3),* pp 349-380.

Hayward, R. J. (2017). Evaluating the imminence of cyber attack for purposes of anticipatory self-defense. *Columbia Law Review, 117(2)*, pp 399-434.

Himes, A., & Kim, B. J. (2021). "Self-Defense on Behalf of Non-State Actors", *U. Pa. J. Int'l L., 43*, 241.

Jain, J. P. (1963). "The Legal Status of Formosa: A Study of British, Chinese and Indian Views", *American Journal of International Law*, 57(1), pp 25-45.

Mačák, K. (2015). Military objectives 2.0: The case for interpreting computer data as objects under international humanitarian law. *Israel Law Review, 48(1)*, pp 55-80.

Schmitt, M. N. (2013). "Below the threshold cyber operations: The countermeasures response option and international law", *Va. J. Int'l L., 54*, 697.

Schmitt, M. (2017). *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (2nd ed.), Cambridge: Cambridge University Press.

Note: Unless otherwise indicated, the term "Taiwan", as it is used in this paper, encompasses the island of Taiwan (Formosa), Penghu (The Pescadores), and other islets which fall under the effective control of the Taiwanese authorities. It is important to note that issues regarding the legal status of Kinmen (Quemoy) and Matsu will not be discussed here.

Digital Forensic Readiness Model for Internet Voting

Edmore Muyambo[1] and Stacey Omeleze Baror[2]
[1]DigiForS Research Group, Dept of Computer Science
[2]University of Pretoria, Pretoria, South Africa

Edmore.muyambo@tuks.co.za,
Stacey.baror@cs.up.ac.za

Abstract: Voting is an exercise of choosing a preferred candidate through a process called an election. In many countries, this exercise is a basic human right. In every election process, there are some pre-requisite processes and procedures which must be set up first. These are essential in the pre-vote-casting stage, during vote-casting and post-vote-casting stage. Electoral disagreements amongst stakeholders and parties of interest are usually experienced in each of the above-mentioned voting process stages. The main points of conflict in an election process are vote rigging and vote fraud. Failure to amicably mitigate these issues can result in a criticised/rejected election result. Therefore, this research aims to address the problem of vote rigging and vote fraud allegations in an election process. The resolution thereof is achieved through the introduction of an online based voting system which is supported by a digital forensic readiness mechanism. Online voting system gives citizens the flexibility to use internet-enabled devices such as cell phones and laptops to cast their votes in a safe, secrete and secure protocol. To address the problem of vote rigging and vote fraud, the online voting system is integrated with cyber security and vote protection mechanisms. The cyber security and vote protection mechanism is based on Blockchain algorithms. A Blockchain-based voting process is a peer-to-peer mechanism where a decentralised database is used to store data. Tokens move directly from one peer (voter) to another peer (candidate). The results are tallied by counting the number of tokens paid to each candidate. Each voter is allocated a Bitcoin token and each candidate is allocated a Bitcoin address. During vote casting, the voter transfers their Bitcoin token into the wallet of a registered candidate. At the end of the voting process, the total number of Bitcoin tokens transferred to each candidate is counted and tallied up. The wallet is loaded with only one Bitcoin token, hence there is no possibility of double voting. The model ensures vote security, anonymity, auditability, accountability, accuracy and uniqueness.

Keywords: Blockchain, Internet Voting, Electoral Data Privacy and Data Security, Cyber Security, Digital Forensic Readiness, Online Voting

1. Introduction

Voting is a process of secretly selecting an option of choice from the available options. The process thereof is called an election. Sovereign states and countries make use of the voting process to elect national leaders i.e. Presidents and/or legislators/parliamentarians.

Each country conforms to its voting mandate as stipulated in its national constitution. For example, In Zimbabwe citizens head to national polls once after every five years. The presidential terms in office are limited to two only (Makumbe, 2014). Whereas in China the presidential term limit has been dealt away with (Jiang et al.,2022). Every voting process follows a pre-set process flow. For example, in a traditional voting process, all legible voters are required to pre-register with the election administration board. The voter registration process verifies the eligibility of the voter for vote casting. The ballot paper-based voting system comes with several loopholes. These include the lack of a fool-proof vote auditing mechanism. These loopholes have been identified to be the main causes of disputed elections, vote rigging and vote fraud allegations. The traditional voting process (ballot paper-based system) can be easily manipulated to facilitate vote rigging and vote fraud as the process involves a lot of manual human activities such as ballot paper printing, ballot paper transportation, manual vote counting and results presentation (Srivastava, Dwivedi and Singh, 2018). When ballot papers are in transit to the command centre for final vote tallying, there is a high risk of getting votes tampered with as human security can be easily compromised. The manual counting process is prone to human error as well as intentional vote manipulation. This opens up a risk of announcing unverifiable and untrusted manually audited votes. Therefore, as a way of mitigating the shortcomings of the traditional (paper-based) voting system; this paper is hereby introducing the Digital Forensic Readiness Model for Internet Voting (DFRMIV). DFRMIV is an online-based technology that can be used to instantaneously flag vote rigging and vote fraud activities in an internet-based voting system. Audit trails and data logs are used to trace electronic vote movement from the polling station to the command centre data centres and the final stage of results presentation. The final results are tallied and reconciled against the votes cast. Election results are expected to perfectly tie back with the votes cast at individual polling stations. If there is any difference picked between the votes cast and the results presented the model would immediately flag a possible vote rigging activity. The DFRMIV would then accurately flag all the numerical differences picked and give a detailed view of the vote deviation.

DFRMIV makes use of some end-to-end encryption algorithms such as Blockchain to protect data integrity. Blockchain preserves vote integrity by its inability to be cracked by hackers and fraudsters. Voting transactions are stored in blocks which are in turn stacked together into a chain of decentralised node ledgers. Data stored on a centralized database through this technology cannot be altered or tampered with. Parties of interest can make use of the DFRMIV's findings to challenge an election outcome. The model's findings show valid scientific proof of any vote-tampering activities which could have happened. This makes the findings eligible for use in a court of law to challenge any vote rigging or vote fraud issues.

Therefore, the problem statement of this paper is that currently there is no digital forensic readiness process that addresses vote rigging and vote fraud issues either in the online voting process or the ballot paper-based process. The main electoral phases vulnerable to vote rigging and vote fraud are, during ballot paper printing, vote casting, ballot paper transportation, results counting and during results presentation.

This paper is structured as follows: Section 2 discusses the background and literature review. Section 3 discusses the proposed model. Section 4 collectively states discussion points. Section 5 presents the conclusion of the research and Section 6 provides a list of references used in this paper.

2. Background

As of now, majority of countries still make use of the ballot paper-based voting system (F.M.Mursi *et al.*, 2013). This means only a small number of countries have adopted the electronic voting system. Most countries reject the internet voting system sighting lack of trust in the system, its complexity and lack of operational skillset.

The ballot paper-based voting system involves a lot of manual interventions. This makes it prone to human errors, mistakes and vote manipulation. This makes the whole system vulnerable to vote rigging and vote fraud. The ballot paper-based voting system used to have the greatest trust threshold during its inception years. However, trust in the system is fast dwindling due to its vulnerability to vote rigging allegations, more so in Africa. In Zimbabwe, during the 2013 election season, the main opposition party (then) Movement for Democratic Change (MDC) claimed that some fraudulently printed ballot papers were used to illegally award the ruling party candidates some rigged votes (Magaisa, 2019). They also claimed that ballot papers used contained some tampered chemical composition which would facilitate vote rigging through a chemical process called paper chromatography. Paper chromatography is a scientific chemical reaction process that dissolves, extracts and precipitate the atomic composition of the subject matter in question (ballot paper in this case) (ROBERTS and WOOD, 1951).

Given the ever-increasing global warming challenges, a paperless based voting system would be a great gesture of embracing the fight against global warming.

Concerns in the online voting system include security issues, privacy, accountability, accuracy confidentiality and trust (Muneer and Shamail, 2013). However, all these concerns can be resolved by the digital forensic readiness framework. A well-integrated online voting system would include a series of interlinked systematic nodes that store, verify, authenticate and present a user-friendly election voting interface. (Matharu, Mishra and Chaudhary, 2014). For a smooth transition, the public would need to be imparted with online voting knowledge. This includes making it clear to voters how the system works and how vote security and confidentiality are preserved by the system in an end-to-end process. Security is a top priority in an electronic voting process (Sharkey and Paynter, 2003). The advantages of introducing a digital forensic readiness model in internet voting included vote fraud prevention, improved vote counting time and results announcement, cost efficiency and reduced number of spoiled ballots (Pawlak and Poniszewska-Marańda, 2019). To gain more community trust in the online voting system, the system must provide vote security and user identity protection. All votes cast must be audited and accounted for. Audit trails must be freely available for all to see. One of the efficient ways to meet the above-mentioned points of concern is through the use of the digital forensic readiness model in the online voting system. The digital forensic readiness model makes use of the Blockchain system to facilitate end-to-end vote security.

2.1 Blockchain System

“Blockchain is a peer-to-peer system based on a decentralised, un-tampered, open and transparent ledger that looks like a read/write database because the data stored cannot be altered” (Rabia, Sara and Gadi, 2021). The technology is popularly known for its robust security mechanism in the digital currency technology called Bitcoin. All transactions and the history of all transactions exchanged between users are stored in a grouping called a

ledger. Blockchain is not controlled by any single authority. It is sharable amongst various users. This allows everyone to check the validity of the chain. (Rabia, Sara and Gadi, 2021). It is visible to all users in it, hence transparency and auditability by any of the users. Blockchain is a safe and secure way of preserving data security. Transactions are stored in blocks which are in turn stacked together into a chain of decentralised node ledgers. Data stored on a centralized database through the means of Blockchain cannot be altered or tampered with. Hence its suitability for use in online voting systems and the digital forensic readiness model.

The Blockchain system is categorized into three groups which are, Public Blockchain (anyone can access and have visibility of transactions in the ledger), Consortium Blockchain (partially decentralised and controlled by more than one organization), Private Blockchain (managed and authorises user participation) (Rabia, Sara and Gadi, 2021).

- The public chain has open access to anyone. Everyone can access, send, receive and authenticate transactions. The mechanism which determines the block to be added to the chain is that everyone can participate and all transactions are open and anonymous. (Chen, Chen and Lin, 2020).
- A Consortium Blockchain is a partially decentralised system that is controlled by more than one organization that can act as a node.(Rabia, Sara and Gadi, 2021).
- A private chain is a completely private Blockchain. All write permissions are only held and managed by the owner. Reading permissions outside the chain may be partially restricted. The whole of the private chain is maintained by all members. This category is characterized by fast transactional speed, low cost, protection of transaction privacy and immutability. However, the degree of private chain centralization is relatively high and it is difficult to achieve consensus trust. (Chen, Chen and Lin, 2020).

Privacy and verifiability are the main aspects of internet voting. Blockchain technology can be semantically integrated with the internet voting process to provide a safe and secure online voting process. Although the Blockchain-based data encryption process is very costly, its high-end security mechanism validates the digital forensic processes.

2.2 Digital Forensic

Digital forensic is a process of answering questions about digital data's current and previous state of events (Shanmugam, 2015). This process can be initiated by digital data attack detection, policy misconduct or unauthorized digital data activities. An investigation can be live (occurs as the system is online) or dead (occurs when the system is offline). Countermeasures can be developed and deployed to eradicate system attacks. These include making use of the Blockchain system to detect data tempering activities and flag them out through authentic audit trails. The digital forensic readiness mechanism provides for an authentic, robust yet transparent vote auditing mechanism. Hence, attempted vote rigging and vote fraud activities would be detected and flagged for all parties of interest to see. Therefore, vote integrity is preserved. Direct beneficiaries of digital forensic readiness for online voting include political parties, independent candidates, electoral boards, election observers, legal services, police force and citizens at large (Gloe, Kirchner and Böhme, 2007). Each group of beneficiaries can access and use digital forensic findings as per their needs. For example, the police force can make use of the digital forensic readiness findings in their investigations to lounge an arrest. The legal services can also make use of the digital forensic readiness findings to lay charges or defend the accused. Election observers can make use of digital forensic findings to give a well-informed report to the international community on how free and fair the whole election process was. Section (3) below gives a detailed view of the proposed digital forensic readiness model. In essence, the model comprises of four main processes. These are; the vote casting process, data storage and the vote reconciliation process. After a voter casts their vote, data gets stored in the electoral board's data store and the digital forensic data store simultaneously. The vote reconciliation process would then run on top of the two data stores to compare and contrast data integrity. Any form of vote tempering would be picked up by the reconciliation process. A report would be generated and presented to stakeholders of the election process for further action.

3. Digital Forensic Readiness Model of Internet Voting

3.1 Introduction

To mitigate issues of vote rigging and vote fraud, this paper introduces the digital forensic readiness model of internet voting. The model is made up of four main layers. These are; the vote casting layer, digital forensic data store, electoral board data store and the vote reconciliation layer.

The main components of the DFRMIV's high-level view are the voting process, electoral board vote processing, digital forensic vote processing and the vote reconciliation process as shown in Figure 1 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 get 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.

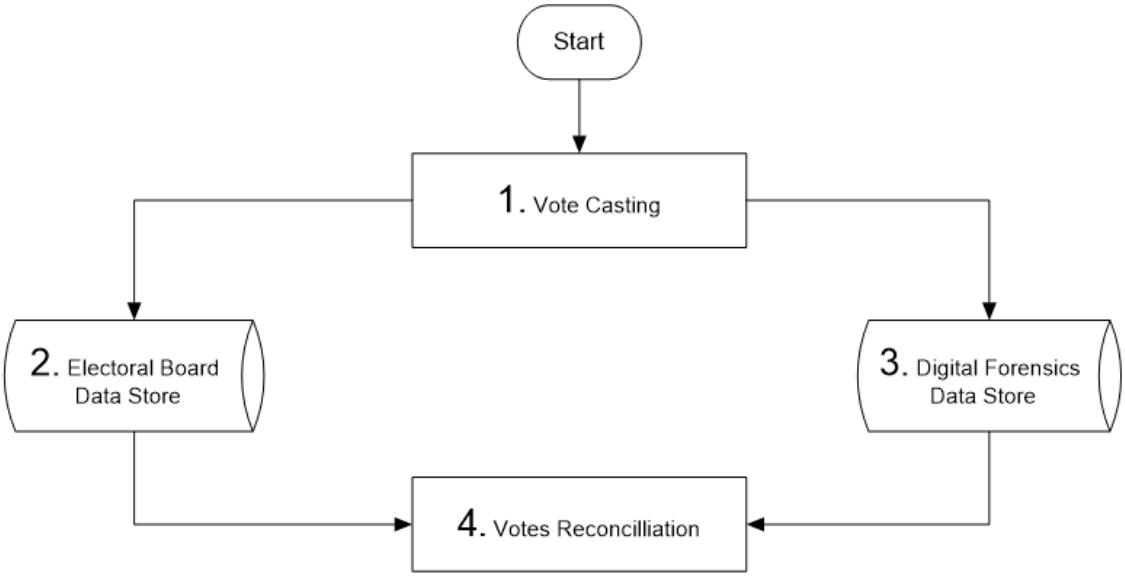


Figure 1: High-Level View of the Digital Forensic Readiness Model for Internet Voting

The Four Main Components of DFRMIV

3.1.1 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 submit the selection. When the submission process is successful the user will be automatically logged out after getting a success status confirmation and the reference number.

3.1.2 Digital Forensic Data Store (DFDS)

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 stores raw data of the votes cast. This dataset is used for digital forensic investigations which include vote rigging detection. DFDS is not government owned, it is owned by independent service providers.

3.1.3 Electoral Board Data Store (EBDS)

Vote's dataset is multicast into the DFDS and the EBDS. This means by default the two data stores should contain the same number of votes. If at one point the two contain a different number of votes, a possible vote rigging activity would have occurred. The EBDS is used by the Electoral Board for vote counting, auditing and results presentation.

3.1.4 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 DFRMIV would automatically detect this and flag all the numeric deviations in detail.

A detailed diagrammatic view of the above four mentioned components of the DFRMIV is shown in Figure 2, Figure 3, Figure 4 and Figure 5 below.

3.2 Vote Casting

The first stage of the voting phase is to identify the voter. After a successful user identification, the system then goes on to verify the user's existence in the voter's roll. If the user's voter registration status gets confirmed successfully they will now be allowed to log into the voting system for vote casting. Once successfully authenticated and logged in, the user can now cast their vote. As soon as the vote is cast, the encryption process (Blockchain) gets triggered automatically. This is one of the most important stages of the digital forensic readiness model as it uniquely identifies each vote and takes account of every vote cast. The system would give a success signal if the vote has been successfully captured and transferred into the secure vote repositories. If a technical error such wrong user name/password or user identification failure the system would direct these glitches to the Help Desk team for further assistance.

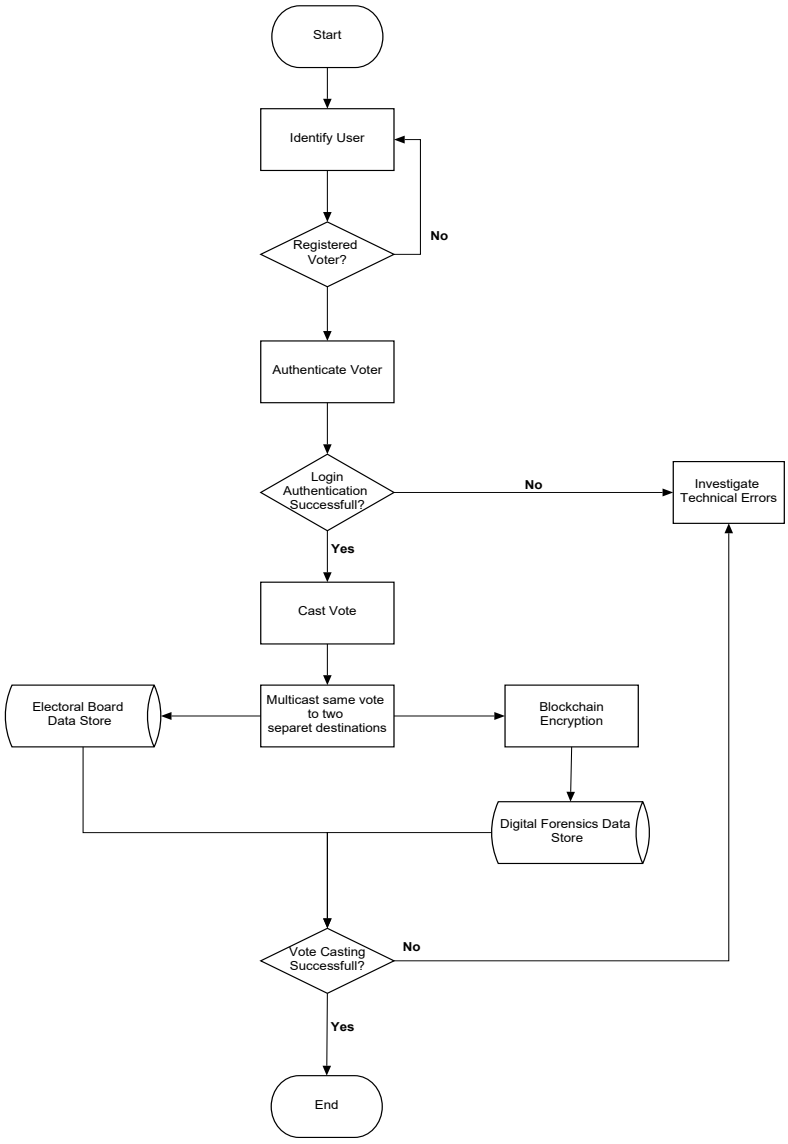


Figure 2: The Vote Casting Process

Components of the Vote Casting Process

Start: The start point of the voting process. The voter gets access to the internet voting device. This can be done either from the comfort of their home or from a polling booth with internet-enabled devices.

Identify User: The user trying to access the voting system gets verified and identified. This is a username and password verification process which is administered by the voting system. This is the first phase of user authentication and it is performed against the system user database. Access to the voting system is granted once the verification and authentication process succeeds.

Registered Voter?: After a successful system log-in, the system checks if the user is a registered voter. This is performed against the voter's roll database. A voter's roll database is managed and maintained by the election administering board.

Authenticate Voter: Once the voter's eligibility for vote casting has been verified and confirmed, the voter gets authenticated for vote casting. This is the second phase of voter authentication. This happens after verifying the user in the system and after verifying that the user is indeed a registered voter.

Voter Authentication Successful?: This is the grand authentication status. If the voter has been successfully identified and verified in all the two stages of voter administration they will then be presented with a voting page. The voting page contains all election candidates as well as the vote casting functionality. If voter authentication fails, the user would be directed to the system's help desk for authentication failure investigations.

Investigate Technical Errors: Voter authentication failures are redirected to the help desk team for further investigations.

Cast Vote: Voter gets to select a candidate of their choice from the presented options on the vote-casting screen. This is achieved by simply clicking on the candidate of choice and clicking the submit button.

Encrypt Vote Data: Soon after vote casting, the vote gets encrypted by Blockchain technology. This is a peer-to-peer process of encrypting the vote and protecting it from illegal manipulation by the way of vote rigging.

Vote Casting Successful: The system check and confirmation whether the vote casting process has completed successfully or not. If yes the process ends right there but if not then the process gets redirected to the system's help desk for vote casting failure investigations.

End: The voting is considered to be complete once the voter has successfully cast their vote. This marks the end of the entire vote casting process.

3.3 Electoral Board Data Store

After a vote has been cast, its next destination is the data repository. In this model, there are two data repositories involved. Firstly the Electoral Board Data Store (EBDS) and secondly the Digital Forensic Readiness Data Store (DFRDS). Votes are stored simultaneously in both repositories as they trickle in from the polling stations. The EBDS is owned by the state. It is used for all the vote administration and processing processes such as vote counting, report presentation and statistical data analysis.

Components of the Electoral Board Data Store Process

Start: Initiation of the vote storage process. After the voter has cast their vote, the voting system takes over the vote transmission process. This is the begging of the process where the vote cast is electronically moved from the front-end (voting application) to the back-end (database).

Cast Vote Transmission: This is an electronic movement of votes from the casting device to the data store repository through a secure internet protocol. Every vote is transmitted safely and securely through the Blockchain encryption process. The encryption mechanism secures the vote and protects it against illegal vote tampering processes such as vote rigging and vote fraud.

Vote Storage: The encrypted votes land into the data store. This can be a SQL Database (On-Premises or Cloud). The data store is also safe from illegal access and illegal data transformation process. The database is equipped with cutting-edge technology that protects the votes cast. Attempted data tampering activities with the intention of vote rigging and vote fraud are picked up and reported in data logs and audit trails.

Vote Storage Successful? A systematic check to ascertain if data has been successfully stored in the repository or not. This check confirms data integrity, anonymity, accountability and availability.

End: This marks the end of the data storage process into the electoral board data store. Successful data storage means the dataset can be accessed by authorized processes only at any time.

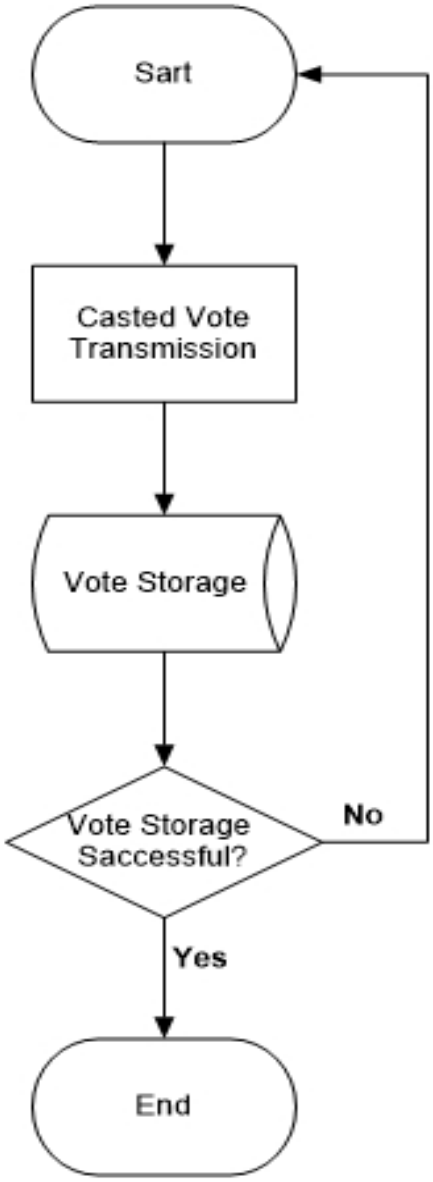


Figure 3: Electoral Board Data Store

The following section explains the Digital Forensic Readiness Data Store process. A detailed data flow process is depicted in Figure 4 below.

3.4 Digital Forensic Data Store

The capturing of data into the digital forensic data store and the electoral board's data store happens simultaneously. The digital forensic data store is not state-owned. The transmitted vote is validated against its Blockchain encryption token. The DFRMIV runs vote tempering checks on top of this data store. This is where the vote rigging attempts are automatically picked and flagged by the DFRMIV system. Data in this repository is verified, safe, authentic and auditable.

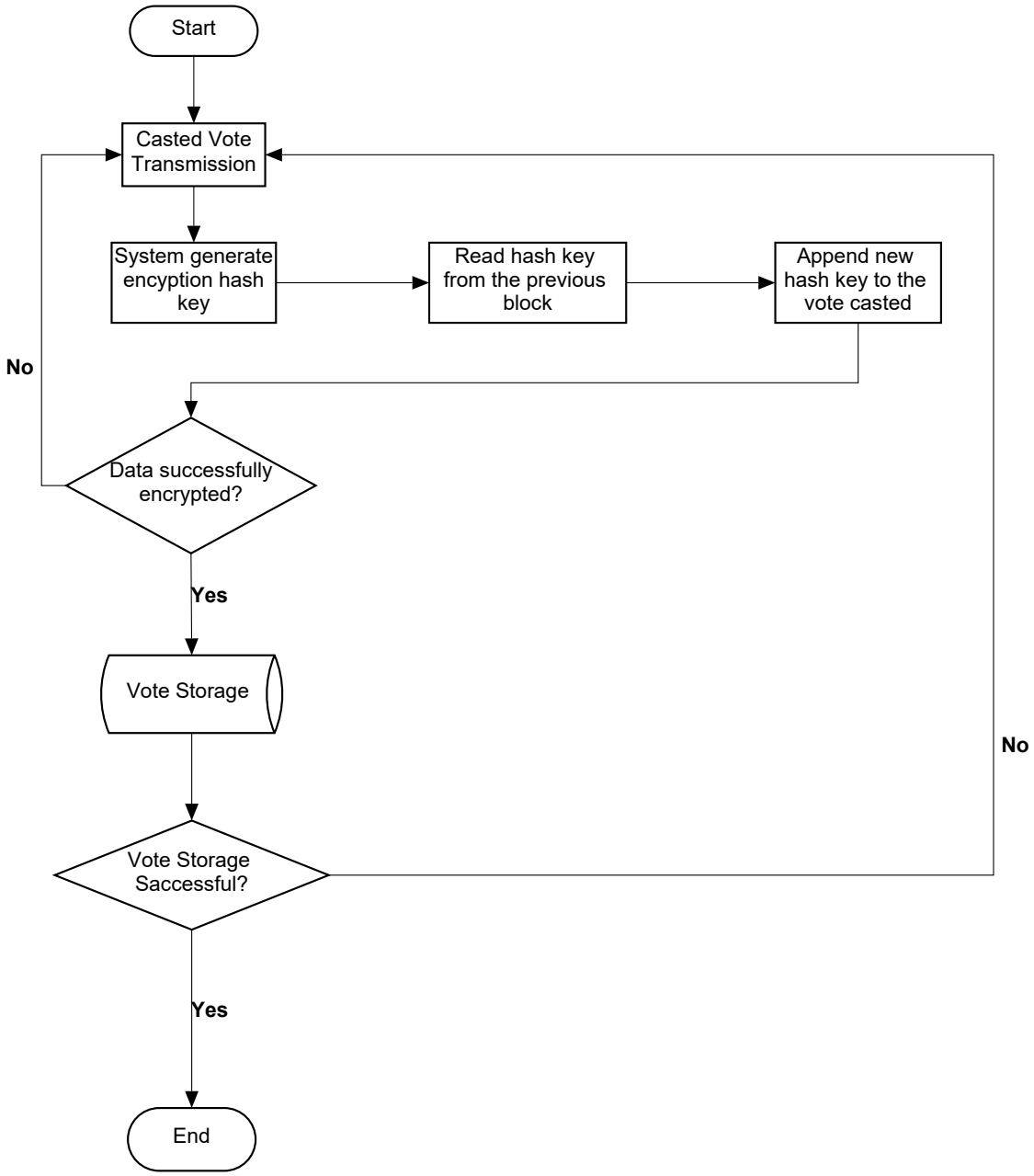


Figure 4: Digital Forensic Readiness Data Store

The stage which follows i.e. the Vote Reconciliation process relies on the Digital Forensic Readiness Data Store to reconcile the election data. Figure 5 below details more on this.

3.5 Vote Reconciliation

The vote reconciliation process references the electoral board's data store and the digital forensic data store to perform a comparison test as a way of vote reconciliation. The Digital Forensic Readiness dataset is considered to be whole, clean and un-tampered with as it went through a series of rigorous vote security mechanisms. A vote rigging flag is raised when the comparison process finds vote mismatches. All vote tempering activities are caught and flagged out at this section of the model. The two datasets are expected to be exactly the same.

A detailed report showing all vote differences and numerical variances is generated by the Digital Forensic Readiness model. Numeric mismatches and data irregularities get logged in audit log trails. The report's contents and audit logs can be referenced by all election stakeholders or beneficiaries. They can also be used in legal processes for vote rigging allegations cases.

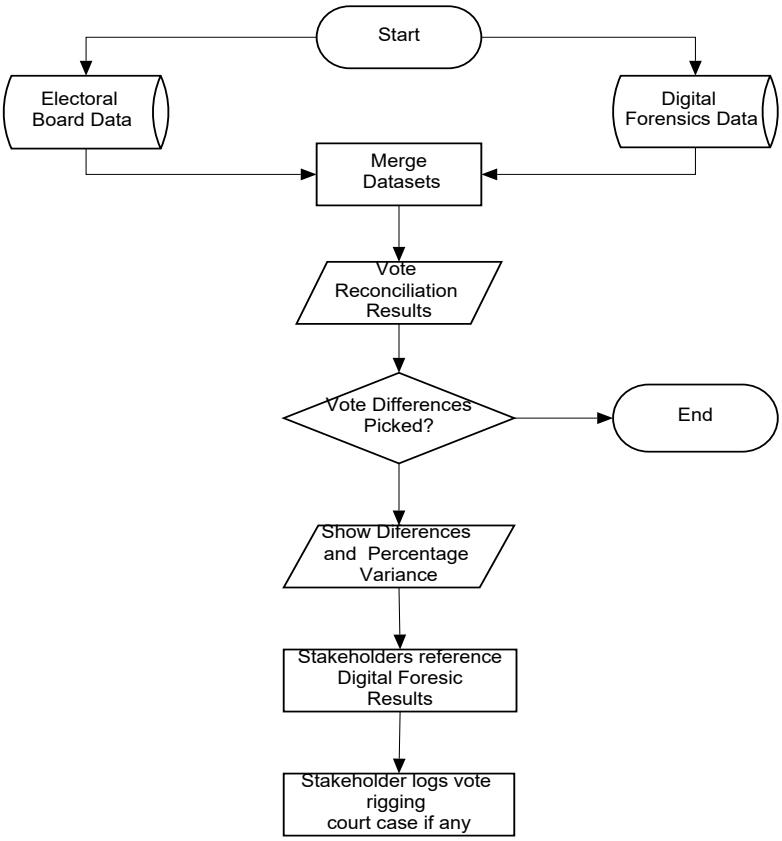


Figure 5: Votes Reconciliation

Figure 6 below shows a holistic view of the entire DFRMIV's integrated components on a lower-level view. The first block encloses the voting process, the second block encloses the data repository mechanism and the 3rd block encloses the vote reconciliation process.

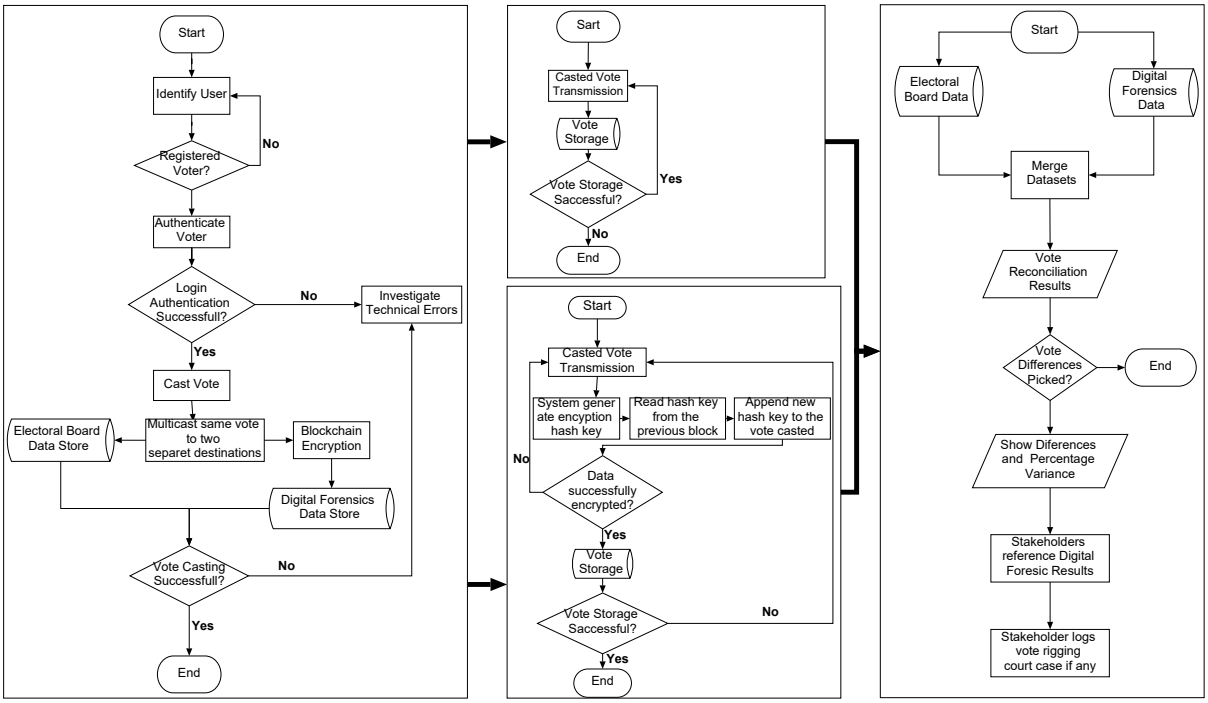


Figure 6: A holistic view of the DFRMIV

4. Discussion

Digital forensic is a scientific data investigation process that makes use of advanced mathematical algorithms to examine data integrity (Thron, Dirnberger and Quirchmayr, no date). Amongst them is the Blockchain technology. Blockchain makes use of advanced security mechanisms to preserve data integrity, security, confidentiality and anonymity. These characteristics make Blockchain the best fit for the development of a tamper-proof digital platform. (Wang, Yang and Hsiao, 2020). Therefore, this technology can be effectively employed in the online voting processes to address issues of electoral fraud, vote rigging and corrupt vote administration issues as it provides a transparent and clear data auditing mechanism. Digital forensic readiness model for internet voting is a scientific concept that is foolproof to vote tampering activities such as vote rigging and vote fraud. When the vote is cast, the Blockchain encryption process executes and protect election data from unauthorised access. This guarantees vote anonymity, security and voter. This is an end-to-end process therefore vote rigging and unauthorised election data tampering is not possible. Therefore, in an event of attempted vote rigging the digital forensic readiness model would flag this in form of digital reports, audit trails and/or log sheets. The digital forensic results can be used in the courts of law to prove alleged vote rigging or vote fraud offenses committed on an election dataset. In this paper, we also looked at the drawbacks of the traditional paper-based voting system. The main causes of concern are vulnerability to vote rigging and vote fraud. Through the means of introducing DFRMIV, this paper outlines how vote rigging and vote fraud can be mitigated in online voting processes. However, since only a fraction of countries thus far have adopted the internet voting system; there is still a long way to go in convincing the international community to trust and employ the DFRMIV to reduce vote rigging problems.

5. Conclusion

The study has shown that a ballot paper voting system is regularly associated with vote disputes. This emanates from loopholes and flaws found in the process, i.e., ballot paper printing, vote casting, ballot paper transportation and storage, vote counting and election result presentation. The audit process in a ballot paper-based system lacks transparency and scientific proof. As a result, the level of trust and confidentiality becomes low. In the ballot paper voting system votes can be easily manipulated and be tampered with without a solid form of trace as this is an end-to-end manual process. The integration of digital forensic mechanism with online voting systems proves to be a better substitute for the traditional ballot paper-based voting system. Its scientific and mathematical semantics make it foolproof to most vote-rigging antics. Its scientific audits can be legally used in law courts to challenge vote fraud and rigging issues.

In the future, this paper's focus is to find ways on how to get the collaboration of online voting and digital forensic to be widely accepted by a greater percentage of the world. The main aim is to eliminate the element of doubt and lack of trust in the online voting model. This would then result in a bigger positive response rate towards digital voting by the international community.

6. Acknowledgments

We would like to thank the anonymous reviewers from the *European Conference on Cyber Warfare and Security (ECCWS)* for their helpful comments. A special acknowledgement to a friend Amie Hewitt for her encouragement throughout the research period. We would also like to acknowledge Job Mbava (brother) for paper reviews at home during the time of writing.

Reference

Chen, M. Te, Chen, C. C. and Lin, T. H. (2020) 'A cryptanalysis of trustworthy electronic voting using adjusted blockchain technology', *ACM International Conference Proceeding Series*, pp. 275–280. doi: 10.1145/3418094.3418143.

Mursi, M.F., Assassa, G.M., Abdelhafez, A. and Samra, K.M.A., 2013. On the development of electronic voting: a survey. *International Journal of Computer Applications*, *61*(16).

Gloe, T., Kirchner, M. and Böhme, R. (2007) 'Can We Trust Digital Image Forensics ?', pp. 78–86.

Kong, H., Kong, H. and Kong, H. (no date) 'Your Meat is My Poison: The Surprisingly Diverging Reactions of Stock Market to the Presidential Term Limit Repeal in China* Junyan Jiang', pp. 1–50.

Magaisa, A. (2019) 'Zimbabwe: An opportunity lost', *Journal of Democracy*, 30(1), pp. 143–157. doi: 10.1353/jod.2019.0011.

Makumbe, J. (2014) 'Elections in Zimbabwe : The ZANU (PF) Hegemony and its Incipient Decline Elections in Zimbabwe : The ZANU (PF) Hegemony and its Incipient Decline', 2(1), pp. 122–139.

Matharu, G. S., Mishra, A. and Chaudhary, L. (2014) 'Integrated election voting system: A model for leveraging ICT in the Indian election scenario', *ACM International Conference Proceeding Series*, 11-16-Nove. doi: 10.1145/2677855.2677944.

Muneer, U. and Shamail, S. (2013) 'Internet voting: A smarter way to vote in Pakistan', *ACM International Conference Proceeding Series*, pp. 348–349. doi: 10.1145/2591888.2591953.

Pawlak, M. and Poniszewska-Marańda, A. (2019) 'Blockchain e-voting system with the use of intelligent agent approach', *ACM International Conference Proceeding Series*, pp. 145–154. doi: 10.1145/3365921.3365927.

Rabia, F., Sara, A. and Gadi, T. (2021) 'A survey on e-voting based on blockchain', *ACM International Conference Proceeding Series*. doi: 10.1145/3454127.3457626.

ROBERTS, E. A. and WOOD, D. J. (1951) 'A study of the polyphenols in tea leaf by paper chromatography.', *The Biochemical journal*, 49(4), pp. 414–422. doi: 10.1042/bj0490414.

Shanmugam, M. (2015) 'Intelligent Digital Forensic Data Analysis and Dynamic Encrypted Distribution Using Global Positioning System'.

Sharkey, E. and Paynter, J. (2003) 'Factors influencing the uptake of online voting in NZ', *Proceedings of CHINZ 2003: The 4th Annual Conference of the ACM Special Interest Group on Computer-Human Interaction New Zealand Chapter*, pp. 121–122. doi: 10.1145/2331829.2331851.

Simons, B. (2012) 'Internet Voting : An Idea Whose Time has Not Come', p. 4503.

Srivastava, G., Dwivedi, A. D. and Singh, R. (2018) 'Crypto-democracy: A decentralised voting scheme using blockchain technology', *ICETE 2018 - Proceedings of the 15th International Joint Conference on e-Business and Telecommunications*, 2(Icete), pp. 508–513. doi: 10.5220/0006881905080513.

Thron, R., Dirnberger, H. and Quirchmayr, G. (no date) 'Requirements and Challenges for Digital Forensic Readiness in Industrial Automation and Control Systems', pp. 232–238.

Wang, P. L., Yang, S. H. and Hsiao, H. C. (2020) 'Hybrid-Voting: A Hybrid Structured Electronic Voting System', *The Web Conference 2020 - Companion of the World Wide Web Conference, WWW 2020*, pp. 83–84. doi: 10.1145/3366424.3382708.

Non Academic Papers

We see what we want to see: Pitfalls of Perception and Decision-making in Security Management

Helvi Salminen
Thales DIS Finland Oy, Vantaa, Finland

helviksalminen@gmail.com

Abstract: We human beings are often convinced of having a clear picture of reality and believe ourselves to be thoroughly rational in our thinking and decision-making. However, our perception of reality is limited and prone to errors, and our decision-making is often guided by emotions and instincts instead of facts and rational thinking. If we don't stop to think we often jump to conclusions based on partial or erroneous information, and eloquently justify our decisions with apparently rational arguments. In many areas of human activities, including security management, limits of perception and errors in decision-making can have harmful, even disastrous consequences. Very often in security management the decision-making process is not sufficiently challenged by critical thinking as decisions are often made hidden behind the veil of secrets. Cognitive biases - systematic errors in thinking affecting decisions and judgments - have been identified and analysed in various contexts, and the results have been applied to improve decision-making processes. However, in the heavily regulated and compliance-dominated world of security management sufficient attention hasn't been paid to cognitive biases and their impacts. As result of insufficient attention an important risk factor is regularly underestimated. This paper includes an introduction to the concept of cognitive biases and the research on the phenomenon. The biases which in the author's experience have a particularly harmful impact on security management are described in detail. This introduction is followed by description of scenarios and real-life examples where erroneous perception and decision-making of security actors leads to disasters. De-biasing is the strategy which aims at eliminating or at least limiting of the impact of cognitive biases. This strategy has been successfully implemented in various types of environments. This paper presents ideas how de-biasing strategies could be implemented in security management in order to improve the quality of decision-making.

Keywords: Cognitive Bias, Security Management, Decision Making

1. Introduction

Cognitive biases are mental patterns which are deviations from rational reasoning. Biases and their impacts are widely discussed in scientific literature and at several popular books and websites. Definitions of nearly 200 biases can be found in various sources.

The Alleydog website (2023) summarizes various definitions of cognitive bias as follows:

- A Cognitive Bias is an involuntary pattern of thinking that produces distorted perceptions of people, surroundings, and situations around us. You can also consider it an altered way of thinking that affects our perceptions and decisions, and can cause mistakes in reasoning, logic, and evaluation.
- Some research suggests that cognitive biases are mental processing "shortcuts" that allow us to make decisions faster when time is a more important issue than accuracy of judgment. These types of cognitive biases are used more frequently when we have limited mental processing capabilities due to lack of time or lack of knowledge about a subject or situation. This is purported to be evolutionary in nature so that we can identify possible dangerous situations quickly.

Cognitive biases are mainly based on *system 1* processes – fast, intuitive judgements and actions without stopping to think.

Biases can be classified into different categories based on their impact and where they apply. They can, for instance impact an individual's decision-making, a group's behaviour, risk taking and motivation. Biases can also cause an individual to misunderstand the importance of alternatives or neglect essential facts in decision-making.

Cognitive biases have also been suggested to be attempts to resolve problems. Why these attempts often fail depends on the complexity of modern society where simple solutions often are not relevant. Valdez et al (2017) present a framework for studying cognitive biases. The article includes cognitive bias codex in which the biases are divided in the following four main categories:

- Too much information –to cope with a flood of information around us the brain filters out most of it, and picks out the pieces which are likely to be useful (for instance a*nchoring* and *confirmation bias*).

- Not enough meaning – the world is not seen as a set of isolated events, instead the brain tries to make sense of it by filling the gaps with already existing knowledge (for instance *illusion of validity* and *impact bias*).
- Need to act fast in a situation of uncertainty it is sometimes necessary to act fast and there is no time to make a profound judgement of all possibilities (for instance *overconfidence*).
- What should we remember – as it is not possible to remember everything one sees and experiences, lots of details are discarded and the remaining information is associated to existing memory structures which further enforces the criteria of selecting what to remember (for instance *misattribution of memory* and *primacy effect*).

Biases are common, and so is the bias blind spot: often people are not aware of their own biases and tend to see other people as more biased than themselves.

Perception error and cognitive biases impact all areas of human activity. This article describes the basic facts of cognitive biases and explains their potential impact on security management.

2. System 1 and system 2

People often prefer to see themselves as thoroughly rational and are convinced to make decisions based on facts. The human mind has the capability of rational thinking and decision-making, but the human cognition has also other characteristics.

According to dual-system theory, human cognition has two distinct types of thinking: one that is fast, automatic and effortless (system 1), and one that is slow, analytic and effortful (system 2). Often an activity is initiated by system 1 but completing it requires the involvement of system 2. System 1 includes the innate skills, system 2 is result of learning.

Kahneman (2012) explains the characteristics of these two systems, and mentions as automatic activities of system 1 for instance the following:

- Detect that one object is more distant than the other
- Orient to the source of a sudden sound
- Detect hostility in a voice
- Understand simple sentences

The following is a list of activities which require system 2 to be activated:

- Focus on the voice of a particular person in a crowded and noisy room.
- Monitor the appropriateness of your behaviour in a social situation.
- Fill out a tax form.
- Check the validity of a complex logical argument.

The influence of system 1 on human reasoning is often underestimated. Kahneman shows that it has an important role in our thinking process. When conscious rational thinking is not guiding the decisions, the result can be error of judgement. These errors are called cognitive biases which are described in the next section.

The advantages and disadvantages of the two systems have been described by Kahneman and several other researchers. The following table shows how Rzepczynski (2023) summarizes the characteristics, advantages and disadvantages of the two systems.

Table 1: System 1 and system 2 characteristics – based on Rzepczynski (2023)

	System 1	System 2
Characteristics	Fast Effortless Unconscious Triggers emotions Associative Looks for patterns Looks for causation Creates stories to explain events	Slow Effortful Conscious Logical Deliberative Can handle abstract concepts
Advantages	Speed of response in a crisis Easy completion of routine and repetitive tasks Creativity through associations, so good for expansive thinking	Allows reflection and consideration of the “bigger picture”, options, pros and cons, consequences Can handle logic, maths, statistics Good for reductive thinking

	System 1	System 2
Disadvantages	Jumps to conclusions Unhelpful emotional responses Can make errors that are not detected and corrected, such as wrong assumptions, poor judgements, false causal links	Slow, so requires time Requires effort and energy, which can lead to decision fatigue

3. Illusions

Many of the judgement errors that manifest themselves as cognitive biases have their roots in incomplete or erroneous perceptions of reality. Chabris and Simons (2009) describe in detail the mechanisms of everyday life perception errors, which they call illusions.

The six illusions having a major impact on people's lives are

- *Illusion of attention* – the belief of processing all information around and failure to understand that observations are limited only to the things on which attention is focused.
- *Illusion of memory* – the misunderstanding of how memory actually works and the overestimating of one's memory capacity and accuracy.
- *Illusion of confidence* – overestimating one's own abilities especially relative to other people, and seeing confidence as evidence of competence.
- *Illusion of knowledge* - a person thinks to know more than he/she really knows.
- *Illusion of cause* – seeing the causal relationship between things where there is none.
- *Illusion of potential* – believing that there are unused hidden potentials in the human brain and these potentials can be released with simple techniques.

The next sections look into more detail about each of the above-mentioned illusions.

3.1 Illusion of attention

According to Saariluoma et al. (2000) thinking starts from stimuli that our senses collect from our environment. Attention can be focused only on one thing at a time, so inevitably many surrounding events remain unnoticed and thus don't have impact on the thinking process. As the information on which decisions are made is incomplete, thinking errors – thinking which produces non-optimal results – are inevitable.

Chabris and Simons (2009) give several examples of in-attentional blindness – a situation in which an individual fails to perceive an unexpected stimulus in plain sight. This failure succeeds when there are numerous stimuli and attending to them all is practically impossible. Attention is given to stimuli which the individual expects to see or which are so salient that they attract the observer's attention.

People who believe their attention covers a bigger portion of reality than it actually does are under the *illusion of attention*. Chabris and Simons have conducted several experiments which verify this claim – people fail to see even a gorilla or a violent beating of a person when their attention is attracted to something else.

The limitations of attention can cause involuntary errors, but they can be used also on purpose to support the objectives of someone who is in control of a situation.

3.2 Illusion of memory

Many people would like to see memory as exact recordings of events. However, the human memory does not work as a video or sound recording device – instead of recording events exactly as they are, the memory makes selections of what to preserve and associates it with already existing memories. Believing to remember more things and more accurately is a symptom of the *illusion of memory.*

Chabris and Simons (2009) also explain how memories depend on the limitations of perceptions. We can only remember things which have been in the focus of our attention, and our attention depends also on what we expect to see. With several examples the authors also demonstrate that memories – even the most vivid ones – change when time passes. Also recalling events from the long-term memory change the memories.

Relying on human memory can have negative consequences if the memories are incomplete or altered – for instance in the case when eyewitness of a crime gives honest but faulty testimony resulting in conviction of an innocent person.

3.3 Illusion of confidence

According to Chabris and Simons (2009) the illusion of confidence has two distinct aspects which are, however, related to each other:

- People tend to overestimate their own capabilities especially relative to other people.
- Confidence is valued as a demonstration of competence.

The first aspect is clearly seen when people driving cars are asked how they consider themselves as drivers. A clear majority claims to be better drivers than the average – which of course cannot be true. Overconfidence – excessive belief in one's superior qualities – can turn against the confident person or company if it leads to lack of self-criticism and failure to see the need for development.

Chabris and Simons state that when starting to learn new skills the original skill level is low and confidence often higher than the reals skills – which means overconfidence. Along with improving skills the gap shrinks. The difference between skill level and confidence is called the confidence gap. Naturally the confidence gap can also be the other way around and the skill level can be higher than confidence in the skills. The following picture illustrates the two different confidence gaps.

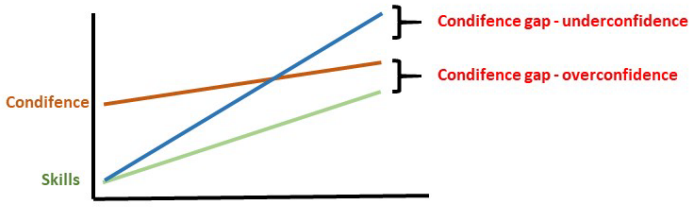


Figure 1: Confidence gap

In groups of humans confident persons tend to dominate communication, and usually are opinion leaders. People can act with confidence for good reasons, but it is not rare that a very confident person manifests overconfidence – not superior competence.

According to Chabris and Simons the most dangerous kind of overconfidence in one's abilities is when we one is unskilled, but the level of confidence is much higher. On the other hand, under-confidence can cause lack of initiative and poor performance.

Harris (2022) describes *confidence cycle* as a model for developing useful skills and achieving the balance of skills and confidence. The cycle is quite similar to the PDCA (Plan, Do, Check,Act) of ISO management system standards.

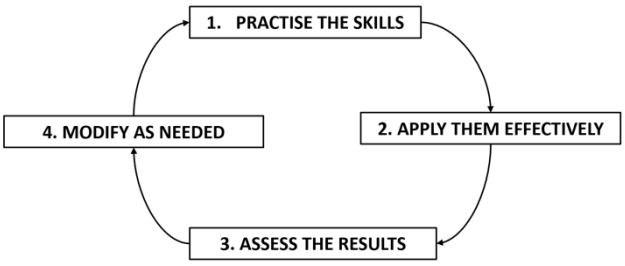


Figure 2: The confidence cycle – based on Harris (2022)

3.4 Illusion of knowledge

When people think to know more than they really do they are under the influence of *illusion of knowledge.* This illusion, like other everyday illusions, is very common.

Even when people know that they don't know everything, they often believe to have the knowledge which in a given context is essential for making rational decisions. The illusion of knowledge combined with the illusion of

confidence in a person is decision-making position is a toxic combination which at worst can cause disasters in organizations and private lives of individuals.

Chabris and Simons (2009) claim that the illusion of knowledge persists partly because people tend to rely on experts who are confident and think to know more than they really two. Some organizational cultures can also favour people who have the combination of illusions of confidence and knowledge.

3.5 Illusion of cause

When people make wrong assumptions about the cause of events, they have fallen into the trap of the *illusion of cause*. This illusion appears in many ways, all of which could be summarized "jumping into conclusions".

One of the factors contributing to the illusion cause is the timing of events. It is quite common for people to believe that earlier events cause later ones even when there is no dependence between the two. Mistaking to think that the correlation of events always indicates causal relationship is quite common.

One form of the illusion of cause are conspiracy theories. A person under this illusion is convinced to have identified in events patterns which are proofs of causal relationships. The observations of conspiracy theorists are guided by "research" in which the result is fixed in advance.

Anyone working in risk management and incident investigation should be very careful not to jump into conclusions. Interpretation of causal relationships should be based on facts, not on opinions.

3.6 Illusion of potential

Many people think that we humans use only 10 per cent of our brain capacity. This idea persists even though there is no evidence to support it. Instead, several sources, including OECD (2023) present several arguments against it, for instance, the following:

- The human brain weighs approximately 2% of the total body weight, it uses 20% of the whole energy. Evolution does not favour waste of resources, so it is unlikely that the sophisticated and complex human brain has evolved to be utilized only with 10 per cent efficiency.
- If we use only 10 per cent of the brain capacity, losing 90 per cent of it should not be a major problem. However, experiences in clinical neurology show that this level of loss has serious consequences.
- The knowledge of functions of different sections of the brain and modern imaging techniques clearly demonstrate that 90 per cent of the brain is not inactive.

So the 10 per cent myth has been busted by several researchers.

Chabris and Simons (2009) describe also another myth – the Mozart effect. This myth states that just by listening Mozart's music the dormant capacities of the human brain are released and can be used. Like the 10 per cent myth, the Mozart effect is not supported by any scientific evidence.

4. Cognitive biases in security management

We have now had a look at basics of cognitive biases. Security management is human activity and thus subject to cognitive biases.

4.1 Security management

The purpose of security management in an organization is to protect organization's valuable assets from threats which could damage or destroy the assets. The assets include people, buildings, machines, systems, business processes, information, reputation – and anything else that the organization needs to continue its activities. The assets must be identified, and the policies and procedures to protect them must be developed and applied.

Several frameworks have been defined for security management either in the form of publicly available standards like ISO/IEC 27001 for information security management or as control frameworks for a specific purpose like Intergraf Certification Requirements for security printing companies. The frameworks define requirements for management procedures or a set of mandatory controls which must be applied in the organization – or a combination of both.

Smith and Brooks (2013) describe in their book about security science three different types of strategic security management frameworks: risk-based, quality assurance, governance and strategic approach. Based on several models components of a consolidated framework of security science is introduced. Topics addressed by the framework include security risk management, business continuity management, technology; physical, personnel and industrial security, investigations, law, criminology, facility management, fire & life safety, and intelligence.

The Confederation of Finnish Industries (EK 2023) has developed a security management model. It is described as nested circles. In the inmost circle there is the target state – business continuity and compliance with security requirements. The target state is achieved with security management which consists of information security, facility management, crisis management, management of fraud and incidents, life safety, personnel security, environmental security, workplace security and security of production and operations.

The scientific approach of Smith and Brooks and the practical approach of the Confederation of Finnish Industries include a lot in common. They both address the same main asset categories and require organization's top management involvement. Many other existing frameworks are also built on same type of layered approach. Strong integration to business objectives is also emphasized I most models.

It is obvious that there is no simple definition of security management. However, keeping in mind the definition that the purpose of security is to protect an organization's valuable assets from different types of threats, it is possible to understand the impact of cognitive biases in managing security. The situation is complicated due to the nature of security – some issues are not open to everyone in the organization, for instance, monitoring methods would lose their efficiency if details are exposed to a large audience. This makes it more difficult to notice if security decisions go wrong due to cognitive biases.

4.2 Impact of cognitive biases on security management

Currently a comprehensive analysis of cognitive biases in security management is not available. This section includes descriptions of some cognitive biases which according to the author's professional experience and discussions with other security professionals could be particularly harmful when appearing in security management. For each bias its potential impact on security management is described briefly in an example case. When the example is from an event reported in public this is indicated. The definitions of biases included in this section are mainly based on Kahneman (2012) and Chabris and Simons (2009).

Anchoring is the tendency to make judgements based on a piece of information (anchor) which can be the first piece of information obtained on a subject, or information recently hear but irrelevant in the given context.

When a person searches and interprets only information which is compatible with existing ideas and discards other information, her/his reasoning is limited by *confirmation bias*.

- The security manager is investigating a major security incident. A previous minor incident was caused by one employee (X) who made a minor deviation from security rules, and was given a formal warning. In the current investigation the security manager concentrates major efforts on the actions of X who is suspended from work. X's actions could not be traced back to be the root cause, but were rather a enabled by some underlying problem. The collected audit trails indicate errors made by another employee (Y), but because this person had a clean record and is a popular person in the work community, the security manager systematically ignores the evidences. Only when a second incident occurs, proper actions are taken.
- In this case the anchor was the previous misconduct of X, and the confirmation bias caused important evidences to be ignored in the investigation.

Framing is a powerful tool in directing people's judgements. By presenting the same information differently causes the receiver to make different interpretations based on the frame more than the actual contents.

- Company X is planning to buy a smaller company. It hires a consultant to make the research of potential targets. The research points out two alternatives. The consultant's report describes A as an innovative company with excellent future prospects, and explains that its poor economic result depends on a major investment in people and innovations. Instead, B which has a stable client base and good economic record, but very little growth in the last couple of years, is evaluated to have lost its innovative power – and for this reason the consultant's recommendation is A.

- The security manager of the company X reviews the risks of the two options. His recommendation is B, because B has a mature management system, which includes well-defined security procedures.
- These two frames – opportunities vs. risks - produce opposite results.

Security frameworks are specific frames. They define topics to which security management should pay attention. However, it is important to maintain an open mind and see if important security topics are not addressed by the frameworks.

Illusory correlation means seeing unrelated events as connected. If we add to the equation the causal relationship between the observed events, the result is the *illusion of cause.*

- Illusion of cause is particularly harmful in security incident investigation and risk analysis. If root causes of incidents are not identified, proper actions to prevent similar incidents in future cannot be taken. The same applies to risk treatment – even if the risk is correctly identified and analysed, failure to understand the causal relationship between the risk and possible consequences can result in inefficient risk treatment.

A person who has *authority bias* gives value to and is influenced by the opinion of an authority, even if there is a reason to doubt.

The following is a real life example (EDRI 2005)

- Sonera (Finnish telecom company which later merged with Telia) management ordered detailed examination of telephone behaviour of employees in order to find out who was leaking information to press. On another occasion the security staff voluntarily and without legal basis provided monitoring data to police to assist in investigation. Five persons got suspend sentences.
- The public sources don't give any clear clues about the impacting cognitive biases. However, considering the persons involved, *authority bias* could be one of them – the actions started from executive management orders which were not challenged by the security team. *Groupthink* (the pressure to agree with the group without criticism) and overconfidence may also have played an important role.

Availability bias makes us overestimate the probability of events and give too much value to ideas which are readily available in mind.

- Many of the issued addressed with security management procedures are known problems and deserve the attention given to them. However, if the security manager fails to "think out of the box" she/he is limited by the *availability bias*, and many important issues which are not inside the standard management frame, or are complex, remain unnoticed.
- The other side of the coin is the *salience bias* which causes us to overestimate the probability and significance of salient events which attract our attention. In the context of security management this could, for instance, occur in case of a minor security audit finding which is erroneously seen as an indication of an underlying major problem. The organization allocates a lot of resources to prevent the same type of minor issues from occurring, and thus fails to identify and correct issues which are more probable and potentially much more harmful.

We fire with a m*ental shotgun* when we answer to questions, and instead of answering the complex question answer a much easier question.

- The security manager is asked to verify if a proposed action is compliant with applicable requirements. The answer is provided quickly – compliance verified – and the actions taken.
- Later, a closer look at the regulations reveals the unpleasant fact that the action would have required approval of competent regulatory authorities.
- *Chauffeur knowledge* - fluently outspoken information without the speaker having real profound knowledge of the topic – is often the factor launching the mental shotgun.

According to Kahneman (2012) psychologists discovered in the 1980's that exposure to a word causes immediate changes in how easily many related words can be evoked. This *priming effect* is not restricted to concepts and words. A person's emotions, thoughts and actions can be influenced by stimuli of which the person is not even aware. The priming effects can also be concatenated and one priming can launch another. As the impacts of

priming don't happen in the conscious mind, they may be even harder to recognize than other types of biases. Effective priming requires knowledge of the culture of the target environment.

- Security frameworks require systematic and documented management of risks. The risk management procedure can be distracted by discussions which emphasize risks which receive lot of public attention but are irrelevant in the target environment. As result organization's resources can be spent on risk treatment plans which have no impact on the risk level, and more critical risks are not addressed.

5. Debiasing – limiting the impact of cognitive biases

5.1 Debiasing techniques

Various techniques, methods and interventions have been designed to limit the impact of cognitive biases. There are several types of de-biasing techniques targeting cognitive processes, motivation or the environment. To improve the decision-making the de-biasing technique can try to change the decision-maker's behaviour or the environment in which decisions are made.

The decision maker's behaviour can be impacted with suitable targeted training. Morewedge et al (2015) have shown that even a single training intervention can have significant de-biasing effects that persist across a variety of contexts affected by the same bias.

In a decision-making situation the organization can require the persons to consider more than one alternative and be trained in generating alternative solutions to problems. In many situations checklists are a good tool which can prevent skipping essential phases of the decision-making process.

Often things can go wrong because there are organizational habits which may have been relevant when formed but can be harmful in the current situation. Breaking the cycle of habits is not easy, but can be done for instance by rewarding the desired behaviour. Additional information can help when erroneous judgement depends on insufficient knowledge.

Training is essential in debiasing. Soll et al (2015) show that even a single well-designed training event can have long-term impacts.

Useful de-biasing techniques are often context-dependent – what works in one situation may be quite useless in a totally different environment.

5.2 Encouraging example of successful de-biasing

There are encouraging examples of successful debiasing. One of them is the case of RWE (prior to 1990 Rheinisch-Westfälisches Elektrizitätswerk AG) de-biasing project described by Günther et al (2017). RWE is a German company operating in the energy sector. The company's economic result had dropped significantly. and the shareholders were demanding answers from management. The situation was analysed, and the result of the analysis was management underperformance and several cognitive biases. Status quo and confirmation biases had led management to think that things were well as they were and the need for changes was not recognized. Overconfidence and excessive optimism also blinded the management from seeing the need for change.

Once the situation had been analysed, the change project was started. Cultural change was key, and before all big change proposals, it was mandatory to list what de-biasing techniques had been applied when preparing the decision proposal. In the decision-making process a devil's advocate was involved, the role being to present doubt and arguments against the proposed course of action. Naturally, key persons were trained to use the de-biasing techniques.

The debiasing change was a success, and management example and participation was a key success factor.

6. Summary

Cognitive biases are errors of human judgement which impact human activities, including security management. The psychological mechanisms forming the basis of these thinking errors have been studied by several scientists. Debiasing techniques which limit the impact of biases have been developed and successfully applied.

However, there is not much information available on the appearing and impact of cognitive biases on security management. As consequence, also the debiasing techniques which would best suit to decrease the impact of biases is missing. Though *security by obscurity* is not considered to be a valid security management approach, to be efficient in preventing harmful events details of control implementations and investigation methods are not exposed to the entire organization. This makes it more difficult to detect the presence of cognitive biases in security management processes, and as consequence also debiasing techniques are not applied.

Security management is activity which impacts the entire organization by restricting risky actions or by acting as an enabler of business opportunities. Having more information about the impact of biases and suitable debiasing methods could significantly improve the quality of security management. This knowledge could also help to develop better and more efficient security frameworks and procedures.

References

Alleydog (2023) Cognitive Bias definition | Psychology Glossary | AlleyDog.com
Chabris, C., Simons, D. (2009) The Invisible Gorilla, Broadway Paperbacks, New York
EDRI (2005) https://edri.org/our-work/edrigramnumber3-11snooping/
EK – Confederation of Finnish Industries (2023) *https://ek.fi/hyotytietoa-yrityksille/yritysturvallisuus/* (February 28th, 2023) http://eprints.cs.univie.ac.at/5258/1/calero-valdez2017framework.pdf
Günther, B., Heiligtag, S:, Webb, A., A case study in combating bias, McKinsey Quarterly
Harris, R (2022) The Confidence Gap, Robinson, London
Iresearchnet psychology website (2023) https://psychology.iresearchnet.com/social-psychology/social-influence/debiasing/ (February 28th, 2023)
Kahneman, D. (2012) Thinking, Fast and Slow, Penguin Books, London
Kannengiesser, U., Gero, J.S. (2019) Design thinking, fast and slow: A framework for Kahneman's dual-system theory in design, Design science, Cambridge
Morewedge, C.K. et al (2015), Debiasing Decisions: Improved Decision Making With a Single Training Intervention, Questrom School of Business at Boston University
OECD (2023) Neuromyth 4 – we only use 10% of our brain, OECD
Rzepczynski, M (2023) https://www.blogger.com/profile/13692706752978475626
Saariluoma, P:, Maartola, I:, Niemi, P. (2000) Ajattelurskit ja kognitiiviset prosessit taloudellisessa toiminnassa - Thinking risks and cognitive processes in economic activity, Tekes, Helsinki
Smith, C.L., Brooks, D.J (2013) Security Science – The Theory and Practice of Security, Elsevier, Amsterdam
Soll, S., Milkman, C., Payne, J.W. (2015) A User's Guide to Debiasing, Wiley
Valdez, A.C. et al (2017) A Framework for Studying Biases in Visualization Research, http://eprints.cs.univie.ac.at/5258/1/calero-valdez2017framework.pdf (February 28th, 2023)

Zero Trust: The Magic Bullet or Devil's Advocate?

Helvi Salminen
Thales DIS Finland Oy, Vantaa, Finland
helviksalminen@gmail.com

Abstract: The concept of Zero trust was first introduced in mid 1990's, and has gradually attracted increasing attention. This approach to building organizations' information system infrastructures has been developed as response to increasing interaction and interconnection of information systems. Along with organizational boundaries have become less clear with the new business models where a business process exceeds the organizational boundaries, also the boundaries of information systems are no longer clear. In this interconnected world the purely perimeter-based security model defining zones of trusted entities inside the perimeter and the untrusted external world outside the perimeter no longer serves the needs of new business models. And the combination of complex technology and sophisticated attack methods it is no longer possible to be sure that all system components and actors inside the perimeter can be trusted. The Zero trust approach brings the sophisticated controls from the perimeter to the entire system. The core idea can be expressed with the four words "never trust, always verify". No system component is by default trusted , and one-time verification is not sufficient – access to a resource must be verified at each connection attempt. Mutual authentication of the communicating parties is in the core of the approach. But does the zero trust approach have unwanted side-effects? The complexity of the system increases when new control layers are built, and system complexity can increase the possibility of configuration errors. Can there be other side-effects as well? The need for trust does not disappear even when the systems are built on the zero trust principles. When studying the zero trust approach the author started thinking what would happen in human interaction and organizational co-operation if they are based on or partly apply the zero trust approach. And the scenarios were quite gloomy. But is this only a nightmare or already at least partly present in our reality? This article describes the zero trust approach and its applicability to technical environments. The second part present scenarios of the impacts which application of zero trust principles could have – or maybe already has - in human communication and organizational relationships.

Keywords: Zero Trust, Perimeter-Based Security, Organizational Culture

1. Zero trust architecture

1.1 Introduction

The objective of the zero trust approach is to protect an organization's valuable resources which are data, systems and applications. Already the name expresses what is the core idea often expressed as *never trust, always verify*. No subject trying to access a resource is by default trusted, and the access is granted with the least privilege per request. Netskope (2021) defines the primary goal of the zero trust approach to be shifting from *trust to verify* to *verify then trust*.

Rose et al (2020) describe in the NIST special publication (later referred as NIST) an abstract definition of zero trust architecture (ZTA). The following picture presents an abstract model of access. A subject requests access to a resource. The policy decision point (PDP) makes the access decision which is enforced in the policy enforcement point (PEP).

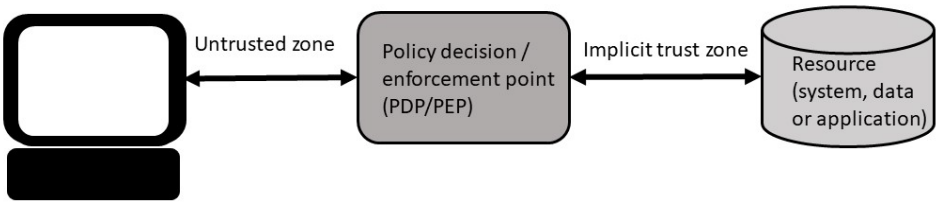


Figure 1: Zero trust access in NIST model

The subjects requesting access are in the untrusted zone, When making the decision the PDP/PEP engine uses a complex set of criteria to evaluate if the requester can be trusted and given access to the implicit trust zone where all entities are considered trusted . The size of the implicit trust zone should be restricted to minimum.

The basic seven tenets of zero trust according to NIST are:

1. All data sources and computing services are considered resources.
2. All communication is secured regardless of network location.

3. Access to individual enterprise resources is granted on per-session basis.
4. Access to resources is determined by dynamic policy—including the observable state of client identity, application/service, and the requesting asset—and may include other behavioural and environmental attributes.
5. The enterprise monitors and measures the integrity and security posture of all owned and associated assets.
6. All resource authentication and authorization are dynamic and strictly enforced before access is allowed.
7. The enterprise collects as much information as possible about the current state of assets, network infrastructure and communications and uses it to improve its security posture.

These tenets define an target situation, and all of them may not be fully implemented in a given zero trust solution. The core components of the zero trust architecture and the trust decision-making mechanisms are described in more detail in section 1.4.

Fernandez (2023) lists a set of fundamental principles that should be applied when building secure systems. These principles which were proposed by Saltzer and Schroeder (1975) include

- *Closed system* – anything not explicitly allowed is forbidden.
- *Least privilege (need to know)* – access is granted only to resources required to perform a task.
- *Complete mediation* – every access request must be validated.
- *Minimal trust* – the parts of a system which must be trusted should be minimal.
- *Holistic security* – all architectural layers must be secure, and security applied during development.
- *Defence in depth* – more than one line of defence is needed.
- *Submarine principle* – some units can be lost, but essential units must be preserved.
- *Compromise recording* – in some case reliable recording of occurred compromise can be used instead of mechanisms which prevent loss.

Many of these principles are a part of zero trust solutions, the exact set is solution dependent. According to Fernandez (2023) all ZTA solutions seem to include closed systems, least privilege and complete mediation.

1.2 A brief history of zero trust

Like many other architectural concepts, zero trust is not a single invention. Instead, today's zero trust is a result of several innovations built on preceding ideas. The following picture presents a timeline.

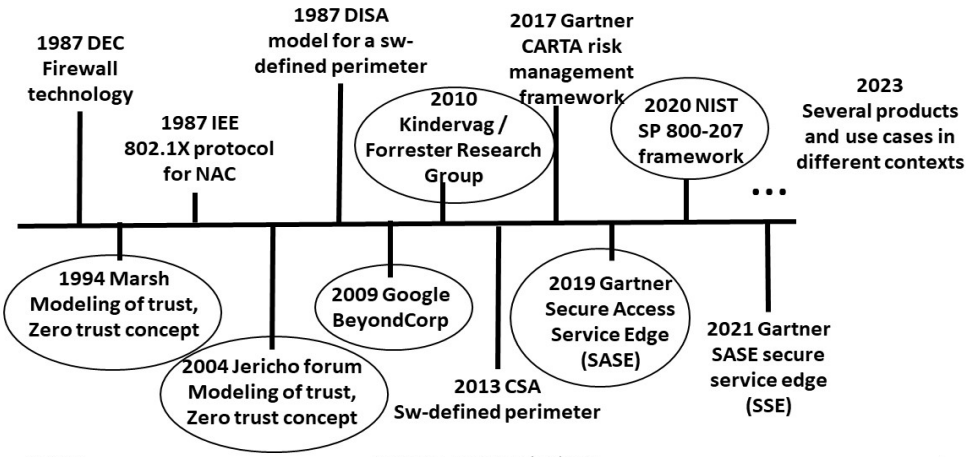


Figure 2: Zero trust timeline – information from several sources

The milestones which are most frequently mentioned in the publicly available zero trust articles and presentations are the following:

- 1987 – Engineers from the Digital Equipment Corporation (DEC) publish the first paper on firewall technology
- 1994 - Marsh (1994) is mentioned in many sources as the first publication to use the term *zero trust*. Marsh's work extends the concept from relationships of humans to artificial agents.
- 2004 - Jericho forum was founded. The mission of the forum was to define and promote de-perimeterization. Core ideas of the forum are presented in the Jericho Forum commandments. The conclusion of the commandments is that the de-perimeterization is inevitable and an organization needs to plan for it and should build a roadmap how to manage the development.
- 2009 – Google BeyondCorp introduces by Ward and Beyer (2009).
- 2010 – Analyst John Kindervag uses the term *zero trust* in a Forrester Research Group paper
- 2019 – Gartner Secure Access Service Edge (SASE) is introduced.
- 2020 – NIST SP 800-207 framework is published.

1.3 Zero trust vs. other architectures

The zero trust approach is often compared with perimeter-based architecture model. The focus of zero trust is in protecting the resources– assets, services, workflows, accounts etc. – regardless of their location. Perimeter-based architecture, instead, protects network segments separated by perimeters with security mechanisms in place. The following picture illustrates the main difference of the two approaches.

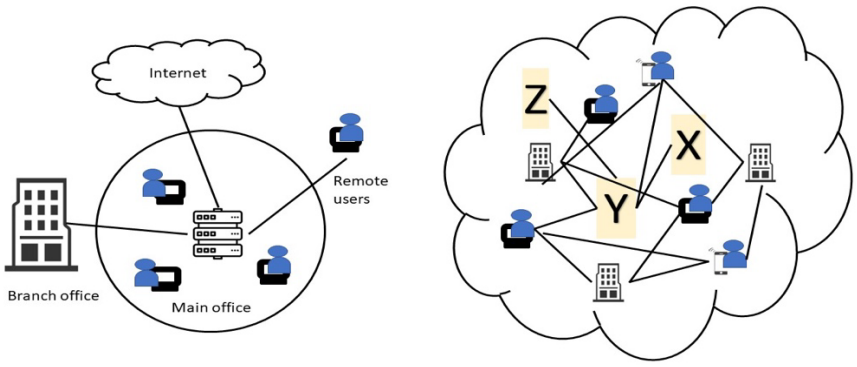


Figure 3: Perimeter-based vs. zero trust architecture

This division is simplified, as also in the perimeter-protected network segments many of the security mechanisms listed in Saltzer and Schroeder (1975) are used.

1.4 Core components of zero trust architecture

The NIST model defines core components which are needed in a zero trust implementation. The following picture presents the components and their roles in the access to a resource.

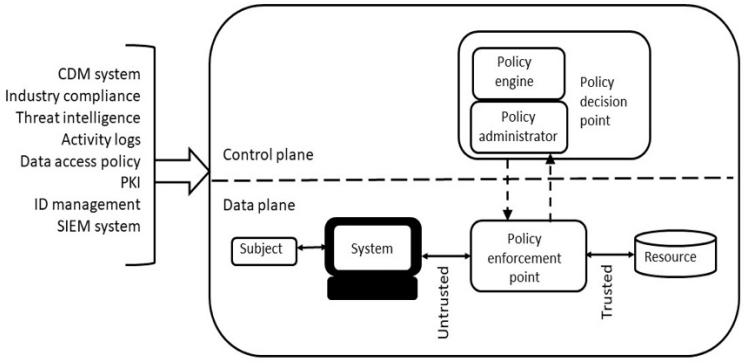


Figure 4: Core components of zero trust architecture in NIST model

- Policy engine (PE) makes the access decision. In the decision it uses the information from the external sources as input to the trust algorithm which makes the decision based on pre-established criteria.
- Policy administrator (PA) allows or denies access based on PE's decision.
- Policy enforcement point (PEP) enables, monitors and in the end terminates the subject's access to the resource. It stands between the untrusted and trusted zones.

Trust algorithm used by the PE is the heart or brain of the architecture. It uses several sources of information when evaluating if the subject can be trusted and given access to the requested resource.

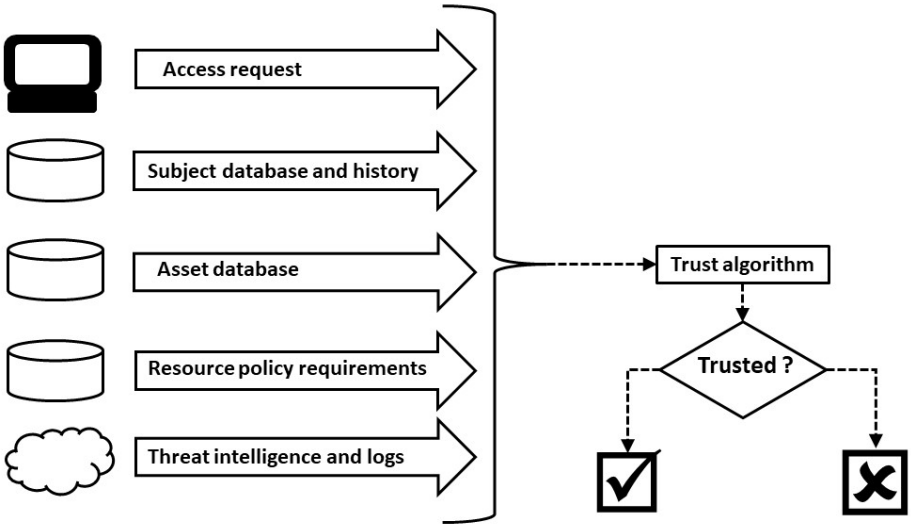


Figure 5: Trust algorithm input in NIST model

There can be different variations of the trust algorithm (TA).

- Criteria-based TA requires all predefined criteria to be met before granting access.
- Score-based TA accepts request which have the trust level above the defined threshold.
- Singular TA evaluates each access request individually and does not use the subject history in the evaluation.
- Contextual TA uses also the subject recent history when evaluating the request.

1.5 Benefits of zero trust approach

Zero trust is not a specific architecture, but a strategy or philosophy to build secure information systems. The main driving force is the change in business models. Organizations concentrate increasingly in their core functions and producing a service requires involvement of several organizations. So, a business process extends beyond the organizational boundaries.

These complex value chains and increasing use of cloud services make it challenging, in many cases impossible, to use purely perimeter-based architectures. Zero trust approach enables protection of the resources regardless of their location. However, physical and network perimeters have not lost their importance in all contexts, which is discussed in section 1.8.

1.6 Potential problems in zero trust approach

As described above, the zero trust approach to information systems architecture embeds security to the complex distributed environment where perimeters no longer can guarantee that all connecting entities can be trusted. However, there are some threats which require attention.
Fernandez (2023) lists potential weaknesses in the ZTA approach. The list includes the following concerns:

- The decision process can be subverted and unauthorized access enabled by attacking the PE or PA.
- Fake component can be introduced to the infrastructure as result of successful attack against the Certification Authority. Fake PE or PA would give the attacker full power to make access decisions.

- Denial of service attacks against the interfaces which handle access requests can block authorized users' access to the system.
- Manipulation or coercion of automated functions, e.g. trust calculations, can cause the system disruption or unauthorized access.

Fernandez et al (2023) also emphasize the importance of governance procedures. This includes the definition of roles, assignment of rights to roles and role and rights management. ZT approach has been criticized for not taking into account the existing risk management governance frameworks or applicable regulations.

In a security architecture the risks of overkill, incompleteness and excess overhead need attention. When evaluating the applicability of ZTA there are some important questions to be asked. Fernandez et al (2023) ask four questions and provide also preliminary answers.

Q1. Does ZTA apply many unnecessary checks? All data items or resources do not need the same level of protection. Equal protection regardless of the sensitivity may be an overkill.

Q2. Does ZTA introduce an intolerable overhead? Dynamic access control and trust evaluation can cause significant overhead, especially if trust must be evaluated separately for each access request.

Q3. Is ZTA implementable for most institutions? ZTA is feasible, but significant support from vendors may be needed in ZTA implementations. Well-planned gradual implementations could be the best strategy, but it must be noted that intermediate stages may not meet the security target of the organization.

Q4. Can ZTA control more, fewer, or similar number of threats as compared with other approaches? ZTA concentrates on the attack surface of the system and makes it more difficult for the attacker to achieve the target. However, ZTA methodology does not ensure that all system levels are covered by the security mechanisms. In this sense ZTA shares the weaknesses of perimeter-based controls.

Townley (2023) points out a misunderstanding about the interpretation what zero trust means – Kindervag's original premise is *trust no packet* instead of *trust no one* as often understood. He also emphasizes the importance of integrating the ZT objectives to the enterprise architecture.

1.7 What remains to be trusted in ZT environment

In the zero trust architecture all access requests of a subject to resources are evaluated based on pre-established criteria. Many implementations also use adaptable criteria which take into account information about recent events. By default the subjects are untrusted.

However, there is at least one essential building block of a zero trust architecture which needs to be trusted: the trust algorithm used by the Policy Engine. The correct configuration and functioning of the trust algorithm is essential to both the resource owner and the requester who needs access to the resource. It is in the interests of both to block unauthorized access and approve legitimate access. Verifying the correct functioning of the trust algorithm is not possible for a basic user, and difficult also for the resource owner in a complex environment.

For a subject accessing a system it is important that the information stored and processed in the system is accurate and up to date. However, often the subject cannot verify this, but has to trust the correct functioning of the system and accuracy and timeliness of the data.

1.8 Physical and network perimeters are still needed

There are many security frameworks which still require information system assets to be placed in secure areas protected with strong physical perimeters – in some cases several nested physical protection layers are required. The same frameworks also include a network architecture with several security layers and strong restrictions to the traffic passing through the perimeters. PCI Council's security standards for card production (PCI-CP) and Finnish National Security Authority's auditing criteria (Katakri) are examples of such frameworks. PCI-CP defines security rules for environments where payment card production data is processed, and Katakri defines security measures of governmental classified data processing.

Physical access to a computer or a network device offers an attacker many interesting possibilities. The attacker can add to a computer a keylogger or USB device which collects data, connect a traffic recording device to a network switch, steal a hard disk from the storage system … and many other things which cannot be done

remotely. Naturally, the attacker can also damage the equipment. Physical security arrangements with several nested perimeters, well managed physical access rights and good physical intrusion detection capabilities give the organization time to react before the attacker has reached the target.

Like in physical security, the *defence in depth* model is in use also in networks where security critical data is stored and processed. In order to reach the target – the valuable data – an external attacker must pass several control points which alert the organization if they detect suspicious activity. This again enables the organization to react to the attack before the "crown jewels" are compromised. Controls must be in place when transferring data between the protected zone and external entities – for instance encryption of data, communication only with pre-approved parties. Naturally controls must be applied also to the entities located inside the perimeter.

Organization's decision on security architecture should be risk-based. The risks to be assessed can be regulatory (compliance with mandatory security criteria), operational, or any type of risk which is relevant for the context.

In a zero trust environment the most severe risk could be compromising of the core components, which make the access decisions. These trusted components should be protected accordingly – which means well controlled physical and logical perimeters.

2. Human aspects of trust

2.1 Trust and distrust

Cambridge dictionary (2023) defines the verb trust "to believe that someone is good and honest and will not harm you, or that something is safe and reliable".

McKnight and Chervany (2000) analyse the challenges in defining trust. They also present a summary of various trust definitions. The following characteristics are included in several definitions of trust: competence, predictability, benevolence, integrity.

The trustor cannot, however, be sure or verify that the trustee meets these expectations. She/he has to interpret the trustee's often indirect signals to make the trusting decision. Bachrarach and Gambetta (2001) describe a trust decision-making model in the form of a trust game. In this game the trustor is in a vulnerable position.

People expect reciprocity in their relationship with others. Hart and Regner (2017) state that lack of reciprocity can create a spiral of distrust in which can be hard to break. Sagarin et al (1998) shed light on the deceiver's position in the trust game. The person who betrays other, assumes that also others are economical with the truth, which gives to the deceiver a justification for not feeling guilty for being dishonest.

Castelfranchi describes trust as a three-level model. First, there must be the disposition to trust. Second, trusting requires trustor's decision to trust, and the final step is the act of trusting.

2.2 Trust in organizational culture

Organizations are what people belonging to the context make them. Klynn (2021) describes the necessity and positive impact of trust in an organization. In a trusting atmosphere we can expect people to be honest and act in a way which is beneficial for both parties. Communication and cooperation are easier when we don't need to doubt the others' words and good intentions. So the mental energy of people can be addressed more to useful and productive work instead of repeated and heavy controls.

Reina (2006) describes a three-dimensional model of trust – transactional trust. The key components of the model are:

- *Contractual trust* implies that there is mutual understanding that people do what they say they will do. Managing mutual expectations, keeping agreements and consistency are essential for contractual trust.
- *Communication trust* main characteristics are the willingness to share information, telling the truth, mutual feedback and speaking with good purpose. Trust and communication are strongly linked with each other.
- *Competence trust* means acknowledging people's skills, allowing people to make decisions.

Lack of trust can have many negative impacts on the organization. If we imagine an organization which has a low level of trust, what could be the impacts of the low trust level in the three components of transactional trust:

- *Low contractual trust* – If you cannot trust that people do what they say and keep the agreements, getting things done and maintaining situational awareness requires significant and detailed monitoring which increases administrative overhead. As people need trusted relationships, low contractual trust could increase the appearance of informal groups. These groups can promote their own agendas, the agendas of different groups may be in conflict with each other and with the organization's objectives. A shadow organization is born.
- *Low communication trust* – When there is low level of trust people tend to withhold information to ensure their own position. On the other hand, having insufficient and inaccurate information is demotivating.
- *Low competence trust* – Competence trust could be described as the *fair play rule*. When people's skills are recognised and they are empowered and can make decisions, the organisational overhead decreases and people see themselves as valuable players in the team. But if the skills are not properly valued and even small decisions must be taken to high level in the organisation people are not committed to the organisation's objectives and can also have a sense of being meaningless and easily replaceable.

3. The impact of technology on humans

3.1 The persuasive power of information technology

A lot of research has been made about human-system interaction (HSI) from different points of view. HSI issues specific to the zero trust environment have so far been addressed mainly as a part of zero trust technical articles – especially the impact of a rigorously tuned trust algorithm which obliges the users to present their credentials frequently, which has a negative impact on the users evaluation of the system.

Information systems are created to serve the business in which they are used. Business application vendors create standard solutions which can be adapted to different use cases. However, a new information system may also oblige to make changes to the business process. So, information systems change the business – not only by increasing efficiency but also by changing business process of the target environment.

But do computers and systems also change the people who use them? And if they do, how?

Ndudule et al (2022) have studied personality-targeted persuasive gamified systems and how effective the persuasive strategies of these systems are in promoting behaviour change. The results of the research show that the strategies can be very effective, but the effectiveness depends on the domain and the personalities of the people involved.

Turkle (2004) describes based on a long term experience the impact that information technology has on people's thinking and mutual relationships. Based on some events she began to study not only what computers do *for* us, but also what they do *to* us. The impact described in the article include the following:

- Transition from slide rules to calculators weakened ability deal with scale.
- People tend to project thoughts and feelings onto their machines, and computational objects are increasingly designed to have emotional and cognitive effects.
- Silent approval of the use of surveillance technology and as consequence more limited (or loss of) privacy.
- Presentation tools like PowerPoint not only provide a way to transmitting content, they also present a model of thinking and aesthetic – and presentation itself becomes its own powerful idea. It is also mentioned that PowerPoint equates bulleting with clear thinking.
- Complex simulations can accustom us to manipulating systems whose core assumptions are hard to understand and may not be true.

Turkle suggests that it is important to develop information technology literacy, and we should be capable of asking the journalists' traditional questions: who, what, when, where, why and how. Since 2004 when the article was written, information technology has increasingly penetrated our everyday lives. And as the tools we use changes us, being aware of these manipulating mechanisms is essential.

Fogg (2003) analyses how computers can have the role of persuasive social actors. A social actors can be persuasive by rewarding with positive feedback, modelling a target behaviour or attitude, and providing social support. And people respond to computer system as though they were social entities. The persuasion uses different types of social cues. The article includes several interesting examples, for instance the following:

- Attractive device or interface has more persuasive power than unattractive technology.
- Computers motivate and persuade more effectively if they show personality traits similar to the user.
- Use of language which indicates social presence stimulates the user to act in a certain manner – for instance to buy more products from an e-commerce site.

A frequent discussion in these days is the impact of social media. These platforms tend to promote powerful emotional expressions which increase polarisation in human relationships and in the society.

3.2 Impact of zero trust architecture on humans

As described in the previous section, the technology itself has an impact on the users. The impact is not limited to the context in which the technology is used, but can extend to the users' perception of the reality.

What specific impacts could the zero trust approach have on its users and on the organizations in which it is used? Based on the information about the impact of other types of systems and software it is possible to create some scenarios.

Let us return to the question how information systems don't only do things *for us* but also do things *to us*. As Turkle (2004) explains information systems don't only transmit content, they can also carry their own way of thinking – the ideas of the creator or publisher of the system. Using the systems can impact our thinking and understanding of the reality in a way that we aren't even aware of.

According to Turkle (2004) PowerPoint has been developed to serve the needs of corporate boardrooms, and the software is designed to convey authority. What kind of thinking could a zero trust system represent? If we ask impacts of zero trust systems from the technical specialists of zero trust could result in the conclusion that the impacts of ZT approach are beneficial – we can trust these systems because they don't trust us as users but verify our credentials by request before granting access. Creating trust by not trusting – a bit of a paradox.

How is the message of the basic assumptions – *never trust, always verify or verify, then trust* – transmitted to the users of ZT systems? Let us have a look at the impact from the point of view of an individual user. If the user does not know the logic of the trust algorithm making the access decision – which often is the case – there is clearly a setup of power distance. The sensation of being powerless is enhanced if the system keeps asking the user to frequently represent the credentials. If people have a repeated experience of being untrusted they also tend to trust others less. The spiral of distrust can fuel itself from this individual experience.

What could the impact be on organizational level? As discussed earlier in section 2.2, low level of trust in an organization has several negative impacts. Lower individual performance and higher level of controls are typical in a low trust environment, which decreases organisation's efficiency. Lack of trust is demotivating, and demotivated people are not likely to take initiative which is beneficial to the organisation.

But do ZT systems increase the level of distrust in the organization more than other types of systems? There is no clear answer to this question, as research results on this topic are not available. Experience with other types of solutions demonstrate that information systems can change the users, and the type of change is linked with the planning premise of the system. ZT systems' basic assumption is *never trust, always verify* – ZT systems could have a negative impact on the trust in the organisation using them. And this can have many harmful consequences, including a spiral of distrust which may be difficult to break. Already the term *zero trust* itself can cause negative associations in people's minds.

4. Conclusion

It is clear that the zero trust approach has been adopted for a good reason – other architectural ideas do not respond to the needs of current business models and complex information systems. When there is a business case, the transition to the new architectural model should be well managed, and risk assessment is a necessary part of the project.

Performance and service level problems are potential issues in a ZT implementation. The systems can also have negative impacts on the user experience and organizational trust. Therefore, a strategy for managing the potential negative impacts of the systems should be part of a ZT implementation project.

It seems that zero trust approach is neither a magic bullet resolving all access problems nor devil's advocate destroying the trust in the organization – it seems to have features of both.

References

Bacharach M., Gambetta D. (2001), *Trust in Signs, Trust in society p 148-182,* Russell Sage Foundation, New York

Bachrach, M., Gambetta, D. (2001) *Trust in signs*, In K. S. Cook (Ed.), Trust in society (pp. 148–184). Russell Sage Foundation

Cambridge dictionary (2023) https://dictionary.cambridge.org/dictionary/english/trust

Castelfranchi, C. (2008) *Trust and Reciprocity: Misunderstandings,* International Review of Economics

Fernandez E.B. et al (2023) *A critical analysis of Zero Trust Architecture (ZTA)* Available in Internet as preprin

Fogg, B. (2003) *Computers as Persuasive Social Actors*, Persuasive Technology: 89-120, Morgan Kaufmall, Burlington

Hart, N., Regner, T. (2017) *The spiral of distrust: (Non-)cooperation in a repeated trust game is predicted by anger and individual differences in negative reciprocity orientation*, International Journal of Psychology, 2017 Vol. 52, No. S1, 18–25

Jericho Forum (2007) *Jericho Forum Commandments*

Klynn, B. (2021) *Building a Culture of Trust,* The Ohio State University, Fisher College of Business

Marsh, S (1994) *Formalising Trust as a Computational Concept,* University of Stirling

McKnight, H., Chervany, N. (2000) *What is Trust? A Conceptual Analysis and an Interdisciplinary Model*, Americas Conference on Information Systems (AMCIS) 2000 Proceedings

National Security Authority of Finland (2020) *Katakri – Information Security Audit Tool for Authorities*

Ndudule, Chinenye et al (2022) *Personality-targeted persuasive gamified systems: exploring the impact of application domain on the effectiveness of behaviour change strategies,* User Modeling and User-Adapted Interaction (2022) 32:165–214, Springer

Netskope (2021) *White paper: Blueprint for Zero Trust in a SASE Architecture, Continuous Adaptive Trust—The Key to Adopting Zero Trust and SASE and How to Get There*

PCI Council 2022) *Payment Card Industry (PCI) Card Production and Provisioning, Logical Security*

PCI Council (2022) *Payment Card Industry (PCI) Card Production and Provisioning, Physical Security*

Reina, D., Reina, M. (2006) *Trust and Betrayal in Workplace,*Berrett-Koehler Publishers Inc, San Francisco

Rose, S. et al (2020) *NIST Special Publication 800-207, Zero Trust Architecture*, NIST Gaithersburg, MD

Sagarin, B., Rhoads, K., Cialdini, R. (1998*) Deceiver's Distrust: Denigration as a Concequence of Undiscovered Deception*, Personality and Social Psychology Bulletin 24(11):1167-1176

Saltzer, J. H.; Schroeder, M. D (1975) *The protection of information in computer systems*. Proceedings of the IEEE, 1975, 63(9).

Townley, Andrew (2023) *Zero Trust: so much noise and confusion* – Newsletter of Archistry Inc

Turkle, S. (2004) *How Computers Change the Way We Think,* The Chronicle of Higher Education

Ward, R., Beyer, B. (2014) *BeyondCorp: A New Approach to Enterprise Security*. Login 20014, Vol. 39, vol. No. 6, pp. 6-11

Zscaler (2022) *A Brief History of Zero Trust*

Work in Progress Papers

Reconnaissance Techniques and Industrial Control System Tactics Knowledge Graph

Eve Cohen[1], Elsa Deitz[1], Jordana Wilkes[1] and Thomas Heverin[2]
[1]The Baldwin School, Bryn Mawr, United States of America
[2]Girls Learn Cyber, Inc., Philadelphia, United States of America

hacker@girlslearncyber.com
thomas@girlslearncyber.com

Abstract: In the initial stages of industrial control system (ICS) penetration testing, pentesters conduct reconnaissance by using various tools including Nmap, Shodan, Maltego, Google, Google Hacking Database (GHDB), Recon-ng and more. Testers use various reconnaissance techniques (RTs) within the tools to directly access ICS devices. Many novice ICS-pentesters stop their reconnaissance work upon successfully accessing an ICS device. However, continuing to conduct reconnaissance after initial access can lead to pentesters finding even more information to find more ICS devices, ICS networks, and ways to make ICS exploitation more effective. Our research motivation stems from finding ways to explicitly model the continuation of using RTs once an ICS device is accessed. Knowledge graphs offer an approach for linking RTs together and creating chains of RTs. MITRE ATT&CK ICS provides a matrix of ICS adversarial behaviours. The matrix consists of main exploit tactics and techniques used to accomplish these tactics. Example techniques include ICS alarm suppression, blocking command messages, starting a device, and stopping services. ATT&CK ICS also provides ICS data sources that defenders use to detect the adversarial techniques. Application logs, files, logon sessions, network traffic, and operational databases represent some of the ICS data sources. We reasoned that if adversaries could find the ICS data sources and discover the ability to modify the data sources, then adversaries could cover their tracks to successfully carry out ICS tactics. For example, ICS attackers could modify log entries to hide the attacker's steps or ICS attackers could delete alarm notifications that showed that ICS attackers changed ICS settings. In this work in progress research, we used knowledge-graph modelling techniques to link together RTs with ICS data sources, the ability to modify the data sources, the ability to then cover tracks of ICS techniques, and the impact of techniques on accomplishing ICS tactics. We named the graph RT-ICS Graph. With knowledge graph queries and shortest-path algorithms run over the RT-ICS graph, we showed how RTs can explicitly lead to impacts on adversaries carrying out ICS tactics. The accomplishment of ICS tactics can cause severe damage or harm.

Keywords: Knowledge Graph, Reconnaissance, Industrial Control Systems

1. Introduction

MITRE ATT&CK Industrial Control System (ICS) (2022) models adversarial behaviours with a matrix of 12 ICS adversarial tactics and 79 techniques used to accomplish the tactics. Example ICS tactics include Execution, Persistence, Lateral Movement, Inhibit Response Function, and Impair Process Control. ICS Techniques include Activate Firmware Update-Mode, Alarm Suppression, Device Restart/Shutdown, Modify Program, and many more. ICS defenders use various data sources to detect ICS Techniques. We reasoned that if attackers could find and modify the data sources, then the attackers could cover the tracks for ICS techniques that are used to carry out ICS tactics. We based this conceptualization on many ICS reconnaissance tasks that we have completed in the real-world. Our preliminary research focused on using a knowledge graph to link the following: reconnaissance techniques (RTs) with ICS data sources, data sources with the ability to modify them, the ability to modify data sources with covering tracks for ICS techniques, and ICS techniques with ICS tactics. The motivation for our research stems from the fact that RTs rarely modelled explicitly even though reconnaissance is a key step in penetration testing.

2. Knowledge Graphs in Cybersecurity

Knowledge graphs have been used in the cybersecurity domain in various ways. For example, Noel et al. (2015) used a knowledge graph to predict the paths of attacks based on attack patterns and the resistance offered by defense technologies. Graphs have also been developed to conduct vulnerability assessments (Ye et al., 2019), security assessments (Zhang and Liu, 2020), and network security analytics (Noel, 2018).

Kurniawan et al. (2022) and Liu et al. (2022), analysed multiple knowledge-graph review studies (Ding, Liu, and Zhu, 2020 and Dong et al., 2020) and determined a key research need: the need to apply graphs to broad, practical issues that organizations face rather than applying graphs to specific network infrastructure instances. We aimed to address this need with preliminary research focused on linking RTs across ATT&CK ICS which is used broadly across all types of ICS networks rather than a particular ICS network. Also, previous cybersecurity

knowledge graph studies have not focused on linking RTs together; therefore, our research offers a novel approach.

3. Knowledge Graph Development and Structure

3.1 Knowledge Graph Development

In domain-focused knowledge graph development, it is critical to rely on domain expertise and reliable data sources (Sattar et al., 2020). The senior researcher of this work in progress paper has over 10 years of ICS cybersecurity experience. Also, all researchers on the team have conducted reconnaissance projects on ICS and successfully accessed ICS devices under bug bounty programs. ICS devices accessed included building automation systems, physical security systems, physical key boxes, and programmable logic controllers (PLCs). Additionally, the team has successfully submitted Google Hacking Database (GHDB) entries focused on ICS devices.

The knowledge graph developed for this research extended a reconnaissance graph called RT-Graph that modelled and linked together over 100 RTs (Deitz et al., 2023). However, RT-Graph contained zero links to ATT&CK ICS concepts. We linked RT-Graph to ATT&CK ICS concepts and named the combined graph "RT-ICS Graph." The foundation of RT-ICS Graph consists of an ontology. Ontologies are used to model domain concepts and relationships among the concepts. When ontologies are filled in with specific data, they turn into knowledge graphs (Kurniawan et al., 2022).

We approached the reconnaissance of ICS as a medium level adversary per the Sandia National Lab Generic Threat Matrix (Woodard, 2007). Table 1 shows the seven attributes from the Generic Threat Matrix used to evaluate threats and the application of the attributes to our ICS reconnaissance approach.

Table 1. Threat Level Attributes Used to Develop RT-ICS Graph.

General Threat Matrix Profile Attribute (Sandia National Lab)	Meaning of Attribute	RT-ICS Graph Approach
Intensity	How far a team is willing to go to achieve objectives; high intensity means the team will risk imprisonment or persecution to achieve goals	Low
Stealth	Level of effort in maintaining secrecy	Low
Time	Amount of time dedicated to reaching objective ranging from days to decades	Low
Technical Personnel	Number of people on the team ranging from individuals to hundreds of people	Low
Cyber Knowledge	Level of network and IT knowledge	Medium
Kinetic Knowledge	Level of ICS knowledge and physical operations knowledge	Medium
Access	Level of ability to place a team member with unlimited access on a device	High

According to the Generic Threat Matrix attributes, our ICS reconnaissance approach focused on using minimal intensity, stealth, time, and number of people to gain the highest level of access. Across various ICS reconnaissance projects, we have repeatedly been able to access ICS devices at the administrator level. We attempted to model our current knowledge in RT-ICS Graph.

3.2 Graph Structure

Ontology-based graphs contain five main entities: classes, objects, object properties, and data properties. Classes represent categories of objects, such as RTs or ICS Techniques or ICS Tactics. Objects are instantiations of classes. Object properties link specific objects together and data properties describe characteristics about the objects. Data properties describe objects. Figure 1 shows the main classes in RT-ICS Graph. ICS Techniques, ICS Tactics and ICS Data Source classes and their objects were drawn directly from ATT&CK ICS. We created the class ICS Ability class to represent the ability of adversaries in modifying data sources found during reconnaissance. Our previous research describes the Reconnaissance Tactic classes in more detail (Deitz et al, 2023).

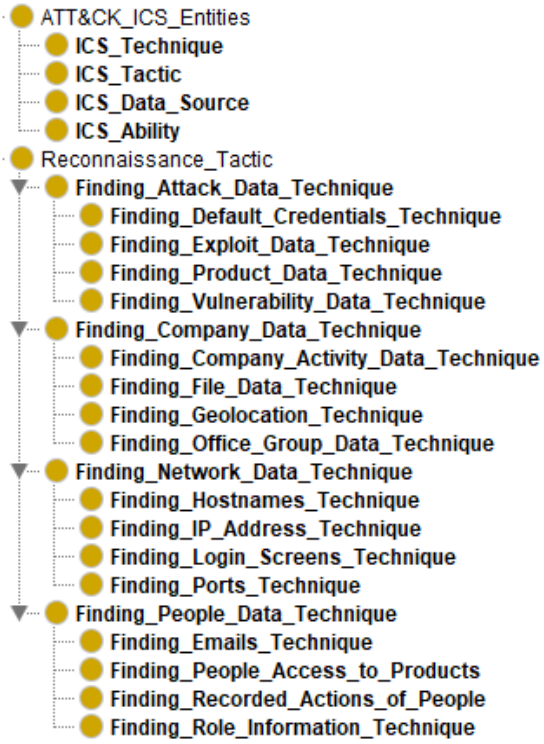


Figure 1. RT-ICS Graph Classes.

Each class is made up of objects which are instantiations of classes. Figure 2 shows objects contained within the ICS Data Source class. According to MITRE (2022), data sources "…represent the various subjects/topics of information that can be collected by sensors/logs. Data sources also include data components, which identify specific properties/values of a data source relevant to detecting a given ATT&CK technique or sub-technique."

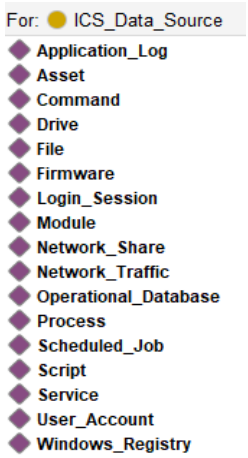


Figure 2. Objects Found in the ICS Data Source Class of RT-ICS Graph.

These data sources come from ATT&CK ICS and were further used to create objects in the ICS Ability class. For example, the data source called "Application Log" was used to create "Ability to Modify Application Log" in the ICS Ability class. This process was repeated for all 17 ICS data sources. The concept of "ability to modify" ICS data sources stems from the researchers' experience in ICS cybersecurity and in ICS reconnaissance projects.

RT-ICS Graph contains several object properties as shown in Table 2. Object properties link together objects. RT-ICS Graph object properties provide direction (one object points to another object) which makes RT-ICS a directed graph. Data properties describe objects. An example data property is ICS Tactic ID such as "T1592" for the ICS Tactic "Gather Victim Host Information." Other example data properties include ICS Technique ID, and Data Source ID all derived from ICS MITRE ATT&CK.

Table 2. Description of Object Properties in RT-ICS Graph.

Domain (Class)	Object Property	Range (Class)	Explanation
Reconnaissance Technique	*providesDataFor*	Reconnaissance Technique	One RT can provide data for another RT; this creates paths of RTs across the graph
Reconnaissance Technique	*findsDataSource*	Data Source	RTs can be used to find various ICS data sources once an ICS device is accessed
Data Source	*discoversAbility*	Ability	Once a data source is found, one can see if the data source can be modified
Ability	*coversTracksFor*	ICS Technique	Once the ability to modify a data source is confirmed, this ability can be used to cover the tracks for ICS techniques
ICS Technique	*accomplishesTactic*	ICS Tactic	ICS techniques are used to accomplish ICS tactics

Figure 3 shows the overall RT-ICS Graph visualized with Gephi, a graph visualization and statistical analysis tool. The Fruchterman-Reingold spatial-layout algorithm was used to produce they graph visual (Fruchterman and Reingold, 1991). Each node on the layout represents an object in RT-ICS Graph (specific instances of RTs, Data Sources, Abilities, ICS Techniques, and ICS Tactics). The arrows represent the object properties from Table 2 specified among specific RT-ICS Graph objects. The purpose of Figure 3 is to provide a representation of how all RT-ICS Graph objects are linked together. Future research shall describe the spatial layout in more detail.

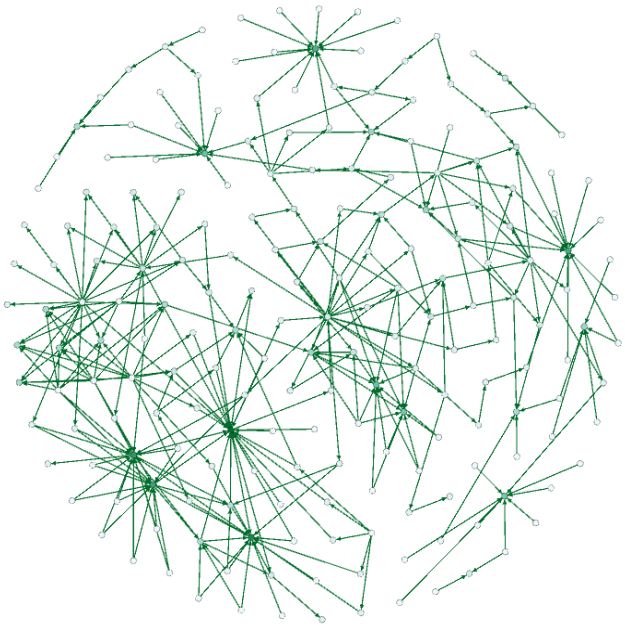


Figure 3. Graph Visualization of RT-ICS Graph Using the Fruchterman-Reingold Spatial-Layout Algorithm in Gephi.

4. Graph Queries and Shortest-Path Algorithm Analyses

4.1 SPARQL for Graph Queries

SPARQL, a recursive acronym for SPARQL Protocol and Resource Description Framework (RDF) Query Language, represents a common query language used to query knowledge graphs (World Wide Web Consortium, 2008). SPARQL holds many similarities with Structured Query Language (SQL). For example, SPARQL uses similar query expressions such as: SELECT, FROM, WHERE, UNION, GROUP, HAVING and more. SPARQL results can provide insights about data and about the connections among data within graphs.

4.2 SPARQL Queries for Preliminary Insights

We developed SPARQL queries to examine how RTs can be used by adversaries once on an ICS device and how these RTs can lead to impacts on ICS Tactics. Across various ICS reconnaissance projects, we have discovered the

ability to modify, upload or download firmware. We developled a SPARQL query focusing on the potential implications of discovering this ability. The SPARQL syntax is as follows:

```
Select    ?ReconTechnique ?DataSource ?Ability ?ICS_Technique ?ICS_Tactic

          WHERE          {?ReconTechnique recon:findsDataSource ?DataSource .
          FILTER (?DataSource = recon:Firmware) .
          ?DataSource recon:discoversAbility          ?Ability .
          ?Ability recon:coversTracksFor ?ICS_Technique .
          FILTER (?ICS_Technique = recon:Module_Firmware)
          ?ICS_Technique recon:accomplishesTactic ?ICS_Tactic . }

ORDER BY ?ICS_Tactic
```

Figure 4 shows the results for the above SPARQL query focused on ICS firmware and shows the ultimate impact on ICS Tactics if adversaries find they can modify the firmware.

ReconTechnique	DataSource	Ability	ICS_Technique	ICS_Tactic
Device_search_for_ability_to_download_files	Firmware	Ability_to_modify_firmware	Module_Firmware	ICS_Impair_Process_Control_Tactic
Device_search_for_files	Firmware	Ability_to_modify_firmware	Module_Firmware	ICS_Impair_Process_Control_Tactic
Device_search_for_ability_to_upload_files	Firmware	Ability_to_modify_firmware	Module_Firmware	ICS_Impair_Process_Control_Tactic
Device_search_for_ability_to_download_files	Firmware	Ability_to_modify_firmware	Module_Firmware	ICS_Persistence_Tactic
Device_search_for_files	Firmware	Ability_to_modify_firmware	Module_Firmware	ICS_Persistence_Tactic
Device_search_for_ability_to_upload_files	Firmware	Ability_to_modify_firmware	Module_Firmware	ICS_Persistence_Tactic

Figure 4. Results of Firmware-Focused SPARQL Query to Show Links from RTs to ICS Tactics in RT-ICS Graph.

One goal of this preliminary research was to show the value of conducting ICS reconaissance tasks even after gaining access to an ICS device. Based on the senior reseacher's 10 years of teaching experience, many novices in reconnaissance stop their tasks once they gain initial access to an ICS device. Novices often lack the insight about the value of continuing to do reconnaissance at this stage; they often miss out on how further reconnaissance can find other critical information to help accomplish ICS tactics. The results of the above query show that RTs can lead to impacts on ICS Tactics that adversaries use to exploit ICS. The above query only focused on firmware. Another example is shown in Figure 5 which is focused on finding the Application Log data source via various RTs and the ability to modify it in terms of user activities.

ReconTechnique	DataSource	Ability	ICS_Technique	ICS_Tactic
Device_search_for_ability_to_upload_files	Application_Log	Ability_to_modify_application_log	User_Execution	ICS_Execution_Tactic
Device_search_for_alarm_status	Application_Log	Ability_to_modify_application_log	User_Execution	ICS_Execution_Tactic
Device_search_for_user_activities_log	Application_Log	Ability_to_modify_application_log	User_Execution	ICS_Execution_Tactic
Device_search_for_ability_to_download_files	Application_Log	Ability_to_modify_application_log	User_Execution	ICS_Execution_Tactic

Figure 5. Results of Application Log SPARQL Query to Show Links from RTs to ICS Tactics in RT-ICS Graph.

In our ICS reconnaissance work, we have found many application logs that track user activities and user accounts. We have found the ability to modify user activities and even the ability to create new users on ICS devices. Figure 6 shows this ability for an Allen-Bradley PLC that we discovered under a bug bounty program. It also shows that we have the ability to create news users with administrator privileges.

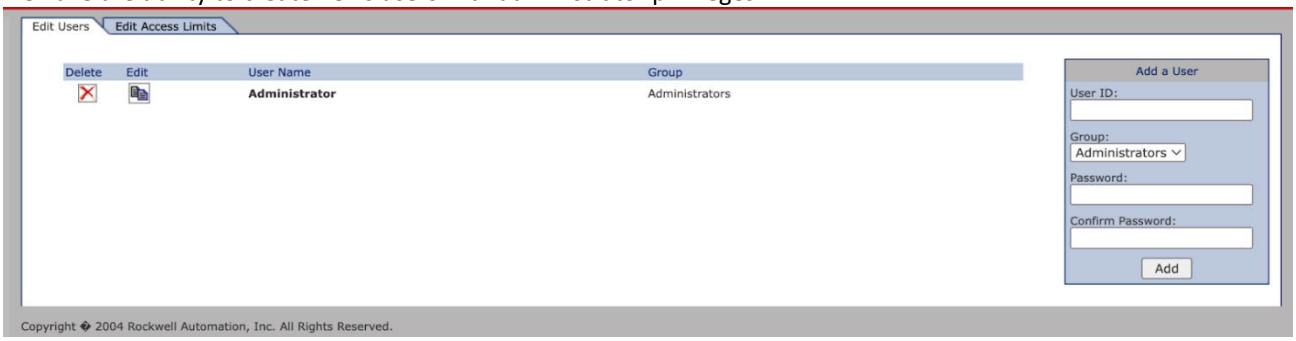


Figure 6. Ability to Modify User Accounts on a PLC.

4.3 Shortest Path Algorithm Analyses

We used Gephi to conduct shortest path analyses from initial RTs to ICS Tactics using the Dijkstra (2022) shortest-path algorithm. For example, we explored various paths from RTs used to scan a hostname with Nmap, a network mapping and port scanning tool, all the way to reaching the ICS Execution Tactic. The ICS Execution Tactic focuses on an adversary running code or manipulating "…system functions, parameters and data in an unauthorized way" severely impacting ICS (MITRE, 2023). Figure 7 shows the shortest path algorithim result. Once the ICS device is accessed via a chain of RTs, starting with a Nmap Quick Scan through a Google search for default credentials to get onto the device, the search for ICS object rules (such as rules for various PLCs) can lead to finding an operational database. If adveraries can modify the operational database, then they can cover their tracks for changing operating modes (an ICS technique) that helps accomplish the ICS Execution Tactic. This is one example of many shortest-path analyses that we ran.

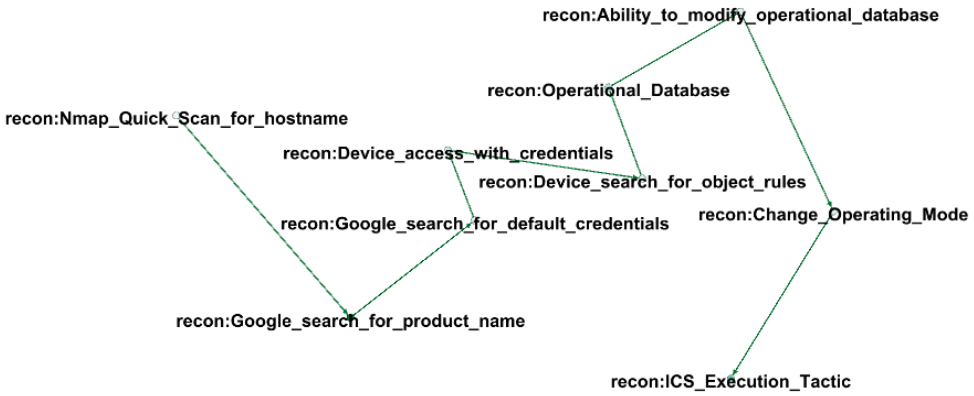


Figure 7. Shortest Path Algorithm Results from Nmap Quick Scan to ICS Execution Tactic in RT-ICS Graph.

Figure 8 shows the shortest path algorithm results starting from an Nmap Intense scan all the way through to the ICS Inhibit Response Function.

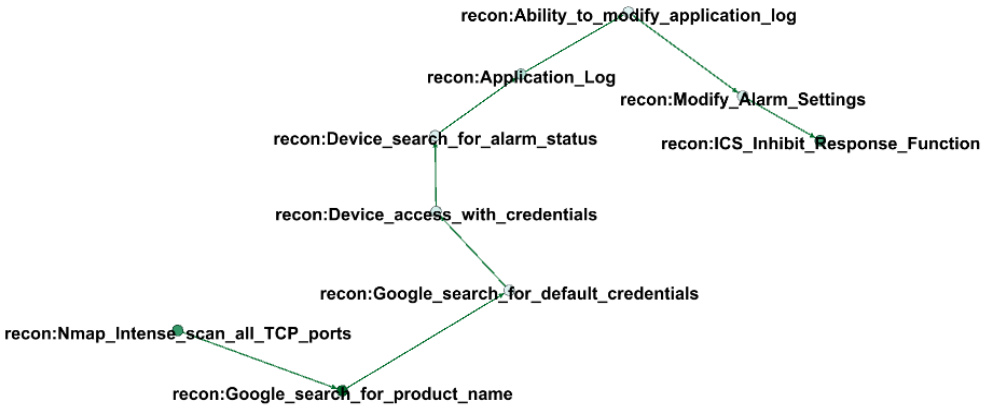


Figure 8. Shortest Path Algorithm Results from Nmap Intense Scan to ICS Inhibit Response Function.

Once an ICS device is accessed, adversaries can search for application logs and the ability to change the logs which then allows the adversaries to cover their tracks for modifying alarms. By modifying alarms, adversaries can better achieve their objective of Inhibiting Response Function. Figure 9 shows the results of bug-bounty reconnaissance work for finding alarm statuses at a university facilities' ICS. Though we did not take actions, one can see that the alarms can be acknowledged and cleared. When adversaries acknowledge and clear alarms, they can distort the awareness of ICS defenders. This can result in adversaries gaining access to ICS systems without being detected and can result in ICS defenders and maintainers lacking awareness of ICS parameters moving towards danger (such as temperatures getting to high, flooding happening, ICS devices overheating) and more.

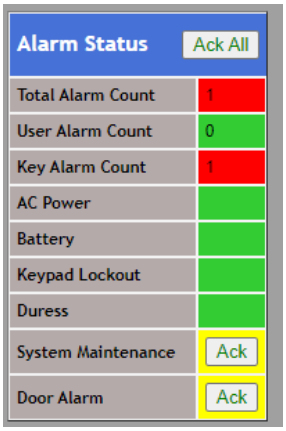


Figure 9. Display Showing the Capability of Modifying ICS Alarms.

5. Conclusion and Future Work

In our work in progress research, we modelled ICS data sources, the ability to modify the data sources, ICS techniques, and ICS tactics used by adversaries to exploit ICS. We connected these ICS concepts to RT-Graph which models over 100 RTs used across various tools including Nmap, Shodan, GHDB, Recon-ng and more. We explored how SPARQL queries can show the value for adversaries in continuing to conduct reconnaissance once an ICS device is accessed. We also explored how shortest path algorithms can show the direct links between initial RTs (such as an Nmap Quick scan) all the way through to ICS Tactics (such as ICS Execution). One limitation within this work in progress research is showing direct impacts on targets. Another limitation is that the modelling does not differentiate between successful execution of steps and when there is not successful execution. Rather, the current modelling only shows potential for successful execution. Future work will focus on adding more RTs to the RT-ICS Graph, conducting more shortest path analyses, and running statistical algorithms over the graph. Finally, ICS experts will examine RT-ICS Graph to provide feedback and improvements on the concepts modelled.

References

Dijkstra, E. W. (2022) "A Note on Two Problems in Connexion with Graphs" In *Edsger Wybe Dijkstra: His Life, Work, and Legacy* (pp. 287-290).

Ding Z., Liu K., Liu B., and Zhu, X. (2020) "Survey of Cyber Security Knowledge Graph*", Journal of Huazhong University of Science and Technology,* Vol 49, No. 7, pp 79-91.

Dong, C., Jiang, B., Lu, Z., Liu, B., Li, N. and Ma, P. (2020) "Knowledge Graph for Cyberspace Security Intelligence: A Survey", *Journal of Cyber Security* Vol 5, pp 56–76.

Fruchterman, T. M. J. and Reingold, E. M. (1991) "Graph Drawing by Force-Directed Placement", *Software: Practice and Experience*, Vol 21, No. 11, pp 1129-1164.

Deitz, E., Cohen, E., Wilkes, J., and Heverin, T. (2023) "Development and Analysis of a Reconnaissance-Technique Knowledge Graph", *18th International Conference on Cyber Warfare and Security*.

Kurniawan, K., Ekelhart, A., Kiesling, E., Quirchmayr, G., and Tjoa, A. M. (2022) "KRYSTAL: Knowledge Graph-Based Framework for Tactical Attack Discovery in Audit Data" *Computers & Security*, Vol 121.

Liu, K., Wang, F., Ding, Z., Liang, S., Yu, Z., and Zhou, Y. (2022) "Recent Progress of Using Knowledge Graph for Cybersecurity", *Electronics*, Vol 11*,* No. 15.

MITRE (2022) "MITRE ATT&CK ICS", viewed January 1, 2023 < https://attack.mitre.org/matrices/ics/>

MITRE (2022) "MITRE ATT&CK Data Sources", viewed January 31, 2023 < https://attack.mitre.org/datasources/>

Noel, S., Harley, E., Tam, K. H., and Gyor, G. (2015) "Big-Data Architecture for Cyber Attack Graphs: Representing Security Relationships in NoSQL Graph Databases", *IEEE Symposium on Technologies for Homeland Security*.

Noel S. (2018) "A Review of Graph Approaches to Network Security Analytics", In *From Database to Cyber Security*, pp 300-323.

Sattar, A., Surin, E. S. M., Ahmad, M. N., Ahmad, M., & Mahmood, A. K. (2020) "Comparative Analysis of Methodologies for Domain Ontology Development: A Systematic Review." *International Journal of Advanced Computer Science and Applications*, Vol 11, No. 5., pp 98-108

Woodard, L., Veitch, C. K., Thomas, S. R., & Duggan, D. P. (2007). *Categorizing threat: building and using a generic threat matrix* (No. SAND2007-5791). Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA (United States).

World Wide Web Consortium (2008). "SPARQL Query Language for RDF", viewed January 1, 2023 <https://www.w3.org/TR/rdf-sparql-query/>

Ye, Z.W., Guo, Y.B., Li, T. and Ju, A.K. (2019) "Extended Attack Graph Generation Method Based on Knowledge Graph", *Computer Science*, Vol 46, No. 12, pp 165-173.

Zhang, K. and Liu, J. (2020) "Review on the Application of Knowledge Graph in Cyber Security Assessment" *IOP Conference Series: Materials Science and Engineering*, Vol. 768.

AI-based Quantum-safe Cybersecurity Automation and Orchestration for edge Intelligence in Future Networks

Aarne Hummelholm, Timo Hämäläinen and Reijo Savola
Faculty of Information Technology, University of Jyväskylä, Finland

aarne.hummelholm@elisanet.fi
timo.t.hamalainen@jyu.fi
reijo.m.savola@jyu.fi

Abstract: The AIQUSEC (AI-based quantum secure cyber security automation and orchestration in the edge intelligence of future networks) brings measurable advances to the cyber security of access and edge networks and their services, as well as Operational Service Technologies (OT). The research aims for significant cybersecurity scalability, efficiency, and effectiveness of operations through improved and enhanced device and sensor securities, security assurance, quantum security, and Artificial Intelligence (AI) based automation solutions. The new application scenarios of near future, the multiple stakeholders within each scenario, and the higher data volumes raise the need for novel cybersecurity solutions. Recently, OT cybersecurity threat landscape has become wider, due to the increase digitalization of services, the increase in virtualization and slicing of networks, as well as the increase in advanced cyber-attacks. Because of recent advances in computing power, AI in cybersecurity analyzing and validations is now becoming a reality. A significant part of currently used encryption technologies which secures communications and infrastructures might become instantly penetrable when quantum computing becomes available. Enabling quantum-safety migration development is a clear goal to the project. The research develops a state-of-the-art information security verification and validation environment that supports the integration of cyber security systems as a reference model, focusing on architectural choices and network connection from different vertical use cases. With the help of the platform and the reference model, common cybersecurity capabilities and requirements can be built, tested, and validated, as well as their fulfillment. In addition to the environment mentioned above, the results of the research are demonstrated and utilized in critical communication systems, water utilities, industrial environments, in physical access solutions and remote work. The developed platform can also be used for auditing devices, systems, and software's in the future. The research integrates new quantum-safe artificial intelligence-based, hardware-hardened, and scalable cybersecurity solutions that have been validated in a standardized way. In this research, we also deal with the requirements of the EU sustainable growth program - issues related to the green transition.

Keywords: edge-intelligence systems, cyber threat, cyber-attacks, AI-based analysing, quantum encryption.

1. Introduction

Our AIQUSEC (AI-based quantum-safe cybersecurity automation and orchestration for edge intelligence in future networks) is research project consortium of Business Finland's Digital Trust program. Cybersecurity is a core part of the program's roadmap. In the research we consider also post-quantum cryptography issues and what challenges quantum computing coming in the use means in our future information's systems and cybersecurity of them (Roger G. Massmann et al., 2023). University of Jyväskylä (JYU) coordinates the project and research. All research partners are from Finland. AI and it use is one part of our research. One of many AI definitions: "The term 'artificial intelligence' means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments (NAII, 2020). The Main Research Question (MRQ) is: Is it possible to build in cybersecurity effectively and efficiently, in a scalable way, for access and edge services in virtualized and sliced integration environment?

1.1 Grouping of paper

Chapter 1 presents the content of the document. Chapter 2 describes future operating environments. Chapter 3 presents our test environment. Chapter 4 describes cyber threats against our information systems. Chapter 5 deals with ethical considerations and chapter 6 international co-operation. Then the conclusions and future work.

1.2 Description of the future operating environment

1.2.1 Threat landscape is getting wider.

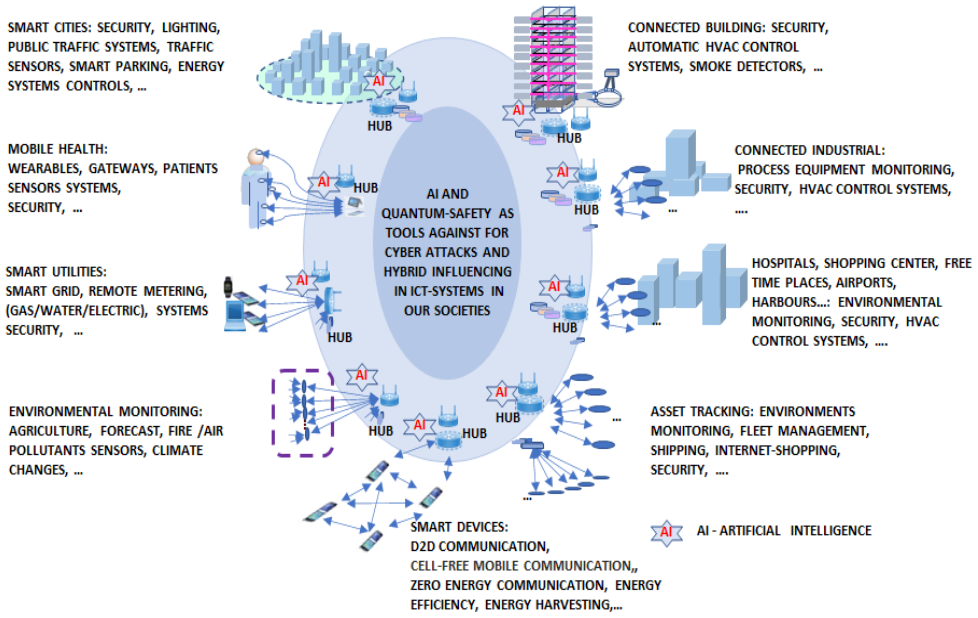


Figure 1. Example of our society's access and edge networks using AI and quantum encryption technologies (Aarne Hummelholm, 2022).

In our society, our devices and services are connected to each other and to other countries. This means a lot of threats and risks, and the situation gives cyber attackers more attack vectors. Threats range from denial-of-service attacks to a variety of new attack methods used everywhere to attack against infrastructures and services across now and in future virtualized and sliced data environments. Figure 1 shows one example of our society's access and edge networks.

1.3 Edge and access services.

Edge computing capabilities and innovations enable value creation and sustainable operations in different vertical use cases, such as smart cities, health, and logistics. Edge computing enhances user data processing in proximity to the point of use the data, introducing data centers functionalities close to the users, for example Multi-Access Edge Computing (MEC) (Hsieh et al., 2020) (ETSI, 2022). Edge access networks are also evolving to include residential and business environments, and mobile networks and virtualization solutions. Rapid development of edge and access technologies calls for more effective cybersecurity solutions than before. The cybersecurity challenges related to rapidly developing future edge and access services are being recognized, and there is a need for novel scalable, effective, and efficient cybersecurity solutions.

1.4 Critical infrastructures and Operational Service Technology (OT).

The services and the critical infrastructures of our digital society should be able to function during and recover from security incidents. Network intrusion by adversaries may lead to a variety of severe consequences from customer information leakage to a cascade of failures, such as massive blackout and destruction of infrastructures. Recent examples of attacks on OT include those suffered by Maersk, Norsk Hydro, SolarWinds, Triton, Stuxnet, as well as power grid attacks in Ukraine. More examples and alarming trends in frequency and sophistication of attacks can be found in the Digital and Cyberspace Policy program's cyber operations tracker – a database of the publicly known state-sponsored incidents that have occurred since 2005 (Cyber Operations Tracker). OT environments are experiencing a major change towards increased cybersecurity awareness (Aziz, A, et al., 2020).

1.5 Quantum safety is a must.

Quantum computing holds great promise for solving some of the most difficult computational problems. But quantum computing also has a dark side. The arrival of quantum computers is imminent and makes arithmetic asymmetric key exchanges unsafe. If data needs to be protected for five to ten years or more, it is necessary to implement quantum secure solutions. It takes up to ten years for countermeasures to be introduced (NICT, 2022). In general, most of the research and development work of the key players is focused on the level of algorithms; while one should think about the whole and that public safety must be considered right from the beginning of the design and development phase, considering the overall architecture and its future use cases. This requires applied research, quantum-safe migration should be possible in different usage scenarios.

1.6 Cybersecurity automation

The application scenarios of edge intelligence will lead to more complexity, with multiple stakeholders, and the growth of collected data, resulting to increased cybersecurity attack surface. Our vision is that especially automation and use of Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) will bring competitive advantage in answering cybersecurity effectiveness, efficiency, and scalability challenges. Because of recent advances in computing power, AI in cybersecurity is now becoming a reality with comparatively small datasets. AI in cybersecurity no longer requires large structures of high-end servers with expensive processors to function, something that was critical just a few years ago.

2. Test and validation systems for cyber threats to the systems of edge intelligence networks of smart societies in the future.

The research aims for significant cybersecurity scalability, efficiency, and effectiveness of operations through improved and enhanced device and sensor securities, security assurance, quantum security, and artificial intelligence (AI)-based automation solutions, test, and simulation environment, figure 2.

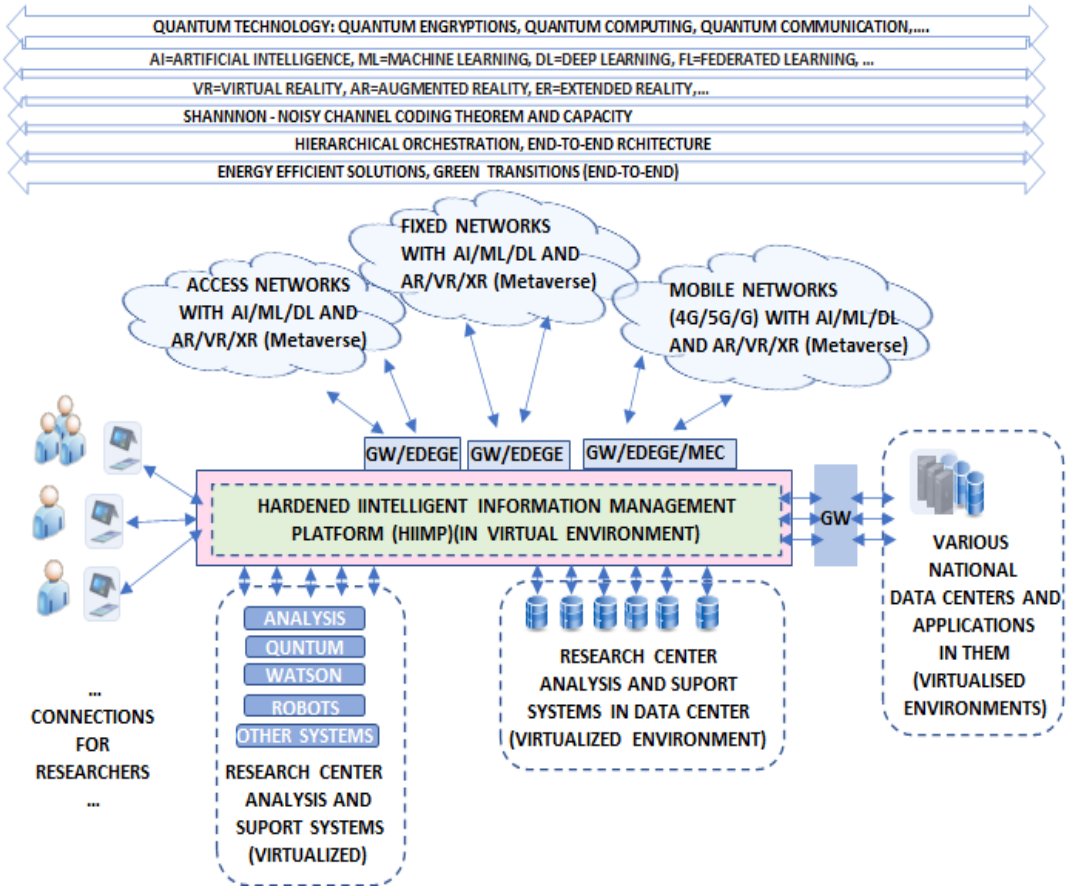


Figure 2. An envisioned security assurance test environment, where cybersecurity system integration reference model is in a core role and together with different simulators (Aarne Hummelholm, 2019).

3. Cyber threats to the systems of edge intelligence networks of smart societies in the future.

AIQUSEC project aims for enhanced device and sensor security, security assurance and artificial intelligence (AI)-based automation solutions, with quantum-safety migration capability. The project will design a leading-edge security assurance and validation environment for edge and access services, with a supporting cybersecurity system integration reference model, with a focus on architectural choices and connection of networks from different vertical use cases. The platform and reference model will enable building, testing, and validating joint cybersecurity capabilities. Attack vectors examples are below.

3.1 Attack vectors in communications networks and devices, figure 3:

- Access networks: Attacks to Mobile End Points, HUB, Smart Devices, IoT devices and sensors (DoS by Flooding, …) (DoS = Denial of Service)
- Attacks on the radio interface (DoS by Jamming) Attack from Physical access to gNodeB (gNodeB = Next Generation Node Base station), …
- Attacks with physical access to the transport Network (Man-in-the-middle attack, Eavesdropping, ...)
- Virtualization: Attacks by Third Party VNF (VNF = Virtual Network function) (Side Channel Attacks)
- Insider Attacks (Data Modification, Data Leakage), API based Attacks (API = Application Programming Interface)
- Attacks from untrusted non-3GPP network (3GPP = The 3rd Generation Partnership Project)
- Attacks from Roaming Network (Theft of Service, Eavesdropping)
- Attacks from Internet and other Networks, (Compromise of Network, DoS, …)

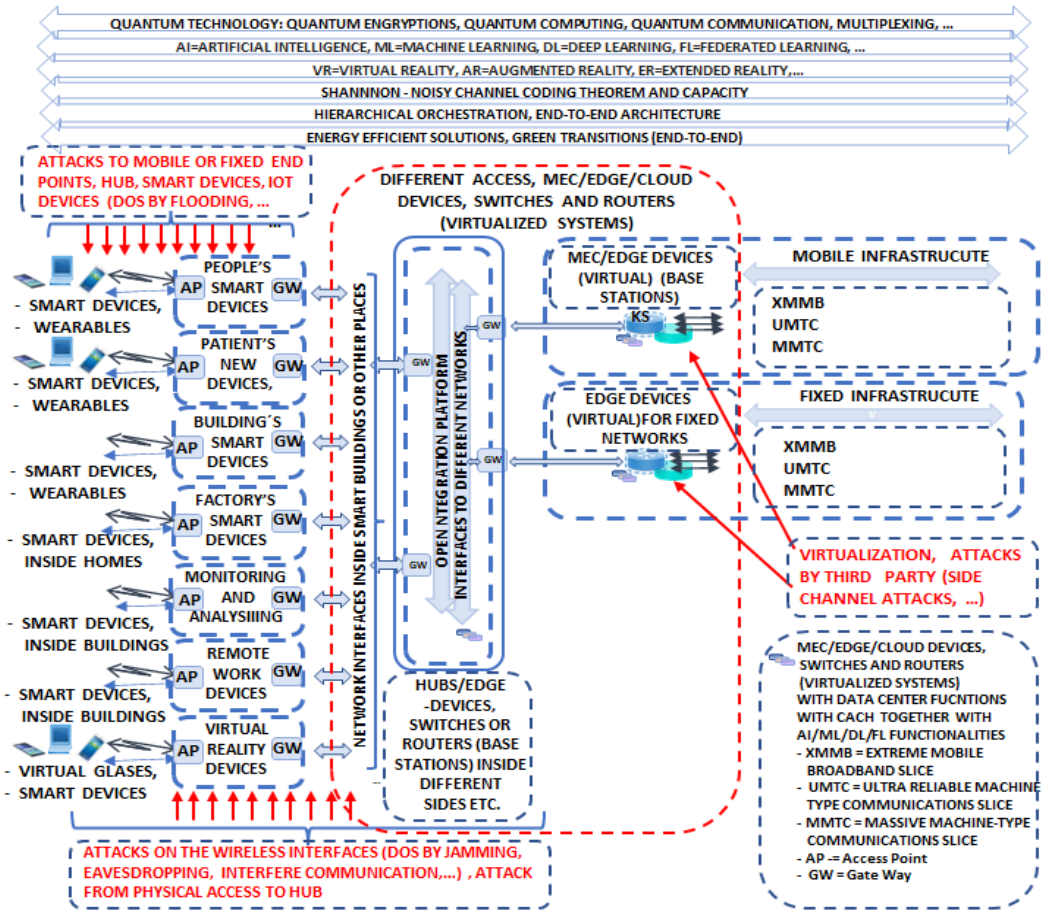


Figure 3. Example of communication systems of future access networks, using artificial intelligence methods and virtual reality interfaces including examples of attack vectors in different use cases (Aarne Hummelholm, 2021), (Ashutosh Dutta, 2021) (Chafika Benza, Tarik Taleb, 2020).

AI/ML/DL-specific threats, figure 3: (Andrew Marshall et al., 2022) (Amit Khullar, Anjani Kumar, 2020) SeukGue Hong et al.,2023)**.** Cyber attackers also use nowadays artificial intelligence methods. Taking advantage of sophisticated and intelligent technology solutions, they can attack everywhere and every system in both the software and physical domains etc.:

- Adversarial Perturbation
    - Targeted misclassification
    - Source/Target misclassification
    - Random misclassification
    - Confidence Reduction
    - Targeted Data Poisoning
    - Indiscriminate Data Poisoning
- Model Inversion Attacks
- Membership Inference Attack
- Model Stealing
- Neural Net Reprogramming
- Adversarial Example in the Physical domain (bits->atoms)
- Malicious ML providers who can recover training data
- Attacking the ML Supply Chain
- Backdoor Machine Learning
- Exploit software dependencies of the ML system
- …

The biggest security threats in using machine learning methods today are data poisoning due to a lack of methods and standards for treats detection and mitigation in this space, combined with reliance on untrusted/unvalidated used public datasets as data sources. Digital Twin and AI/ML/DL are used together to research, developing and optimization of different systems, processes, and services (Cem Dilmegani, 2023). These use cases must be considered also in our work.

AR/VR/XR environments threats, figure 3: (Jassim Happa et al., 2019) (Ben Dickson, 2018). AR (Augmented reality), VR (Virtual Reality) and XR (Extended Reality) devices and systems are used even more in different segments and functionalities of our smart societies. That means that cyber attackers will find ways to exploit vulnerabilities in virtual and augmented reality technology and will use them to infect devices and steal personal data and other types of sensitive data stored in the users' devices. It important to understand these vulnerabilities and find solutions for them. These are the most sensitive aspects which are identified so far:

- Ransomware Risks
- Fast Releases - Security Checks are maybe late
- Gaining Control of Devices
- Interference with AR Medical Devices
- Sabotaging VR/AR Shopping Apps…
- Social Engineering Attacks and must
- …

Metaverse is new working environment, which uses AR/VR/XR devices and systems, and it must consider also in our research and test environments (Metaverse, 2018).

4. Ethical considerations

- When AI is used in decision support automation, it needs to be interpretable, in order the general legitimacy and lack of public trust not to become the showstoppers of advanced automation. That would be unfortunate, as on the other hand AI automation could improve safety and security, by its accelerating effects to rapid response (to, e.g., detection of hazards to human lives in an ongoing mission). The project work studies AI capabilities that can be made transparent to the public safety agencies.

- Research integrity: The European Code of Conduct for Research Integrity (European Code of Conduct for Research Integrity of ALLEA) (All European Academies) and ESF (European Science Foundation) of March 2011 are followed.
- Experimentation subjects: N/A, no persons/animals involved.
- EU directives (EU GDPR, EU MDR, ...) must be considered and fulfilled.

5. International co-operation

International co-operation will be carried out daily throughout the project, in research publications and technical issues in co-operative networks. Specific planned co-operation includes also:

- Research exchanges (1) from JYU - University of Jyväskylä to NR – Norwegian Computing Centre (Oslo), concentrating on the cybersecurity system integration techniques and security metrics and measurements. (2) from JYU to Swiss company ID Quantique SA (Geneva), focusing on quantum-safety.
- Standardization of relevant parts of the reference model is planned to be opened as a standardization proposal. Direct collaboration with ISO/IEC JTC 1/SC 27 is planned. Moreover, the project activities link to IoT and digital twin standardization. Direct collaboration with ISO/IEC JTC 1/SSC 24 is planned. Arto Toppinen of Savonia is SESKO member and a project editor in several ISO/IEC SC41 groups related to IoT.
- Connect research cooperation between Savonia and JYU to iTrust – Centre for Research in Cyber Security (Singapore University of Technology and Design) in the critical infrastructure sector.

6. Conclusions

Answer to the main research question (MRQ): The requirements and fulfillment of a cyber-secure and hardened test and validation environment suitable for testing and analyzing various smart devices, sensors and systems are currently very challenging because information and communication infrastructures are changing rapidly.

In virtualized and sliced information and communication infrastructures services will be orchestrated end-to-end, which means that the same infrastructure's resources will be used and shared by different telecom operators, service providers, private operators, application providers, data center resource providers, service providers including many very critical areas service. These areas will be challenging for this project.

Many different smart devices, sensors, actuators, IoT devices, etc. are connected to our different information environments (pictures 1 - 3), and the devices' hardware and software are in different stages of their life cycle. This is a demanding and big challenge area in the whole project.

The big question is whether these companies participating in the research project have time to get sufficiently extensive information about future operating environments, their development, and challenges in this project time.

Enabling quantum-safe migration development is a clear goal of the research. This area of quantum technology is also one big challenge, because our project is quite short and is there enough time to develop devices and applications, considering the future cyber security challenges caused by quantum computing and AI?

This situation is a big new challenge for research partners, what to include now and what to leave for the next research phase.

7. Future work

One of the project's important future works is international cooperation, which is carried out daily throughout the project in research publications and technical issues in cooperation networks. The specially planned cooperation also includes research exchange, peer countries are selected for active research and business cooperation in the form of joint workshops and visits. The partners also conduct research in many highly critical areas in the field of telecommunications, cyber security, industrial automation, Smart City environments, the water sector and physical access solutions. In standardization issues, the standardization of the relevant parts of the reference model will be opened as a standardization proposal.

One of the project’s future works will be integrating new quantum-secure AI-based, hardware-based, and scalable cybersecurity solutions validated in a standardized way.

One of the project's future research works will be Metaverse and Mixed reality (MX) research.

In this research, we also must consider the coming EU directives that deal with artificial intelligence and its use in our operating environments (EU, Brussels, 21.4.2021).

Reference

Aarne Hummelholm, "Future Smart Societies' Infrastructures and services in the Cyber Environments", Springer, 2022, p. 151. – 182.

Aarne Hummelholm, modified from: "E-health systems in digital environments", ECCWS 2019.

Aarne Hummelholm, modified from "Cyber Security Analysis for Ships in Remote Pilotage Environment", ECCWS 2021.

Amit Khullar, Anjani Kumar, INFOSYS, Knowledge Institute," AI and ML in Cybersecurity Risk Management", Whitepaper, 2020, Available from: https://www.infosys.com/iki/perspectives/cybersecurity-risk-management.html

Andrew Marshall et al., Jugal Parikh, Emre Kiciman and Ram Shankar Siva Kumar," Threat Modelling AI/ML Systems and Dependencies", 11/02/2022, Available from: https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml

Ashutosh Dutta, "5G Threat Vectors", IEEE, Future Network, Artificial Intelligence, and Machine Learning, 27-29.9.2021

Aziz, A.; Schelén, O.; Bodin, U., "A Study on Industrial IoT for the Mining Industry: Synthesized Architecture and Open Research Directions". IoT 2020, 1, 529-550, Available from: https://doi.org/10.3390/iot1020029

Ben Dickson, "Will AR and VR create new cybersecurity threats", February 6, 2018, Available from: https://bdtechtalks.com//2018/02/06/ar-vr-security-privacy-concerns/

Chafika Benza, Tarik Taleb, "AI for Beyond 5G Networks: A Cyber-Security Defense or Offense Enabler", 2020, Aalto University, University of Oulu, Finland

Cem Dilmegani, "Digital twins in 2023: What it is, Why it matters & Top Use Cases", AI-Multiple, January 2, 2023, Available from: https://research.aimultiple.com/digital-twins/

Cyber Operations Tracker is provided by Council on Foreign Relations, Available from: https://www.cfr.org/cyber-operations/#Glossary

ETSI, MEC security; Status of standards support and future evolutions, 2022, Available from: https://www.etsi.org/images/files/ETSIWhitePapers/ETSI-WP-46-2nd-Ed-MEC-security.pdf

EU, "LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS", Brussels, 21.4.2021, COM (2021) 206 final

EU-GDPR, The General Data Protection Regulation, 2016/679.

EU- MDR, The Medical Devices Regulation, 5/2017.

Hsieh et al., Liu, Guo, Gazda, Task Management for Cooperative Mobile Edge Computing, 2020 IEEE/ACM Symposium on edge computing (SEC), Available from: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=928585

Jassim Happa et al., Mashhuda Clencross, Anthohy Steed, "Cyber Security Threats and Challenges in Collaborative Mixed-Reality", 09 April 2019,University of Oxford and London, Available from: https://www.frontiersin.org/articles/10.3389/fict.2019.00005/full

Metaverse v.3.0, 2018, Available from: https://mvs.org/learn/.

NAII, About Artificial Intelligence, 2020, DIVISION E, SEC. 5001, Available from: https://www.ai.gov/about/#NAIAC_-_National_AI_Advisory_Committee

New Technologies– "Tempting Target for Cyber Criminals", Available from: https://arpost.co/category/technology/virtual-reality/

NICT (National Institute of Information and Communications Technology), "Overview of the functional structure of Beyond 5G/6G to achieve the SDGs and realize Society", 5.0, 18.02.2022,

Parachute Technology: "2022 Cyber-attack statistics, data and trends", Available from: https://parachutetechs.com/2022-cyber-attack-statistics-data-and-trends/

Roger G. Massmann et al., Nick M. Grantham, Akalanka B. Mailewa," Quantum Computing: An Assessment into the Impacts of Post-Quantum Cryptography", Saint Cloud State University, St. Cloud, Minnesota, 2023, Available from: https://www.researchgate.net/publication/369776982

SeukGue Hong et al., HyungBin Seo, MyungKeun Yoon, "Data Auditing for Intelligent Network Security Monitoring", IEEE Communications Magazine, March 2023

Water/wastewater utilities leveraging IIoT, Available from: https://iiot-world.com/connected-industry/water-wastewater-utilities-leveraging-iiot/

Security Issues of GPUs and FPGAs for AI-powered near & far Edge Services

Stylianos Koumoutzelis, George Avdikos, Titos Georgoulakis, Ioannis Giannoulakis and Emmanouil Kafetzakis
EIGHT BELLS Ltd, Athens, Greece

stelios.koumoutzelis@8bellsresearch.com
george.avdikos@8bellsresearch.com
titos.georgoulakis@8bellsresearch.com
giannoul@8bellsresearch.com
mkafetz@8bellsresearch.com

Abstract: Graphics Processing Units (GPUs) and Field Programmable Gate Arrays (FPGAs) are widely applied to cloud and embedded applications in which such devices are applied to near and far edge computing operations. This pool of available devices has a wide range of power/size specifications to support servers ranging from big data centres to small cloudlets, or even down to embedded systems and IoT boards. Overall, the most prominent devices and vendors in the market today are the following Xilinx for FPGA-based accelerators, Nvidia and AMD for GPUs, Intel for FPGA- /GPU-based accelerators. Decreasing the latency and increasing the throughput of Artificial Intelligence Functions (AIF), either for network automation or user applications, requires some sort of parallelization inside such purpose-built hardware acceleration. The AI@EDGE project is developing a Connect-Compute Platform (CCP) in which hardware accelerators (1 Nvidia GPU Tesla V100 (near edge device) and 1 Jetson AGX and 1 Jetson Nano (far edge devices), as well as 2 Xilinx FPGAs Alveo U280+U200 (near edge devices) and 1 Versal VCK190 and 2 Zynq ZCU104) are placed inside a server node and execute edge computing scenarios involving multiple nodes of diverse compute capabilities each, to test various integration approaches, to study orchestration techniques measure AIF deployment efficiency, all while developing certain FPGA/GPU code to accelerate representative AIFs of AI@EDGE. In this paper we compare the power/size/performance specifications of all accelerators and highlight the security issues associated with the cloud and embedded accelerators. This study presents the security issues announced by the vendors with the results of our tests and proposes tests and security functions (policies and objectives) which will be applied to the CCP to increase the security level of CCP. It also considers security issues related with the hardware set-up (accelerators inside server nodes) from the network point of view.

Keywords: GPUs, FPGAs, AIFs, AI/ML, Connect-Compute Platform

1. Introduction

The use of hardware accelerators, such as GPUs, FPGAs, TPUs, and VPUs, has become increasingly prevalent in various applications, especially in cloud and edge computing (*INACCEL* et al. 2018). These devices are designed to perform complex computational tasks, process large amounts of data, and support machine learning algorithms. However, as the reliance on these devices grows, so does the need for addressing the security concerns associated with their use. In many cases, these devices handle sensitive information and critical applications, making them vulnerable to various types of attacks and security breaches. To ensure the safe and secure use of hardware accelerators, it is essential to understand and mitigate the potential security risks involved. This paper examines the security concerns associated with modern applications of GPU edge acceleration, including unauthorized access to sensitive data, tampering with the device or its firmware, and exploitation of vulnerabilities in the software or hardware. We also explore the current best practices and recommendations for mitigating these security risks and ensuring the secure deployment and use of hardware accelerators in modern applications.

2. Embedded systems in edge architectures

Recently, embedded systems have emerged as a popular solution for providing low-latency, real-time processing capabilities for a wide range of tasks, making them particularly useful in edge computing architectures and a preferable choice for many modern applications of GPU and FPGA edge acceleration. GPU and FPGA-based embedded systems are being deployed in a wide range of applications, including IoT devices, edge gateways, and mobile devices, supporting data collection, data processing, and control functions.
Embedded systems of such kind, despite their widespread usage and utility, are vulnerable to various security threats and breaches as they often handle large amounts of sensitive data and perform complex machine learning tasks (Pottie & Kaiser et al). This could possibly make them targets for attackers who may seek to gain unauthorized access to sensitive data or tamper with the system's firmware, exploiting vulnerabilities in the

system's software or hardware. Therefore, it is essential to implement robust security measures in order to ensure the secure deployment and use of GPU and FPGA-based embedded systems in edge computing (Ghodke et al., 2021). This could include the implementation of encryption and authentication techniques to secure sensitive data and prevent unauthorized access, the use of firewalls and intrusion detection systems to prevent attacks, and the conduct of regular security audits to identify and address potential vulnerabilities.

3. Securing Embedded Accelerators

As discussed in the previous section, the increasing reliance on hardware accelerators to process large amounts of sensitive data and critical applications, has made these devices potential targets for security breaches. This section delves into the risks associated with unauthorized access and tampering with hardware accelerators and their firmware.

Specifically, tampering with hardware accelerators can occur through the use of malware, spyware, or other malicious software that infects the device and compromises its functionality. Attackers may also resort to physical tampering, such as through the use of rogue devices or unauthorized modifications, to gain access to sensitive data or compromise the device's firmware. The consequences of tampering can be severe, potentially leading to the failure of the entire system and exposing sensitive data to potential misuse.

In addition to malware and physical tampering, attackers may also exploit vulnerabilities in the software or firmware of hardware accelerators to gain unauthorized access to sensitive information. These vulnerabilities can range from simple programming errors to more sophisticated attacks that exploit weaknesses in the system's design. For example, side-channel attacks can exploit variations in power consumption or electromagnetic radiation emitted by the device to extract sensitive data.

To mitigate the security risks associated with tampering with hardware accelerators, it is essential to implement robust security measures. This includes regular updates to software and firmware, as well as the use of secure boot mechanisms that verify the integrity of firmware before it is loaded. Additionally, physical security measures, such as tamper-evident seals and secure storage, can help prevent unauthorized access to the device.

4. Exploitation of vulnerabilities in software and hardware-based attacks

Hardware accelerator weaknesses can be popular soft spots for malicious actors in order to initiate attacks through software-based methods. Although these attackers cannot implement new functions through techniques such as buffer overflow or code injection, they can manipulate inputs, including but not limited to configuration parameters or memory values, hence taking advantage of design flaws. For instance, they may exploit weaknesses (Joshi & Jain et al.) in the accelerator controller to execute various attacks, or target vulnerabilities in the System on a Chip (SoC) design.

Another vulnerability commonly found in hardware accelerators lies on the system bus, as a third party could potentially manipulate it in a way that allows them to execute unauthorized operations on the accelerators. If the system lacks proper security measures, this could result in compromised execution and possibly lead to access to sensitive information either through the output port or by exploiting the shared memory space. Thus, it is important to have proper verification and protection in place to ensure the secure execution and outcome of the accelerators.

Even though the accelerator's specifications may be secure, the actual hardware implementation can still be hacked or generally exploited through physical attacks (Pottie & Kaiser et al). These potentially malicious actions exploit the weaknesses of the implementation and can therefore cause critical failure of protection mechanisms. Side-channel attacks are used to steal secret information from embedded devices and cloud servers, such as in the case of Advanced Encryption Standard (AES) algorithm, which is mathematically secure but vulnerable to physical attacks like power and timing attacks. To counteract side-channel attacks, accelerators can help improve security, for example, by providing constant execution time to make timing attacks ineffective. Intel has already added AES-NI instructions as an example (Ghodke et al., 2021). Accelerators need to be protected from various attacks, including fault-based and side-channel attacks. If not protected properly, a circuit separate from the rest of the processor can become an easy target for precise power side-channel attacks, resulting in the compromise of key information.

5. Proposed testing and security functions

The focus here is on hardware installations which are applied at near and far edge computing applications which run AI/ML operations that require heavy computational workload. A test rig for the purposes of AI@EDGE project involved the deployment of servers at Kernel-based Virtual Machines (Ubuntu-Server) with the CPU/RAM resources specified above and a Kubernetes cluster that consists of the aforementioned VMs and is able to manage the additional custom resources, more specifically a Jetson AGX (GPU) first, and an Alveo U280 (FPGA) at a second stage either for near or far edge applications. Verification process and safety under attacks can be tested and potentially ensured if the virtual machines keep the guest operating systems patched and protect their environment. One can consider disabling unnecessary functionalities in order to minimize the use of the virtual machine console, thus conforming to modern guidelines used as best practices by several actors in the hardware acceleration community.

Especially in the case of FPGAs, the process of their programming consists of loading a bitstream file into the FPGA. This bitstream file contains the binary sequence for the FPGA design. These files are usually vulnerable to the same security threats as regular software including malware, unauthorized copy and Intellectual Property (IP) theft. Cyber criminals can duplicate or reverse engineer FPGA applications by reading the internal memory, intercepting the bitstream, or removing the protective cover of the chip. To a favorable end (Wang, Zhang, & Li et al.), application developers do not reveal their design to FPGA vendors or suppliers. As a result, attackers cannot discover any application-related information by attacking the FPGA vendor. In addition, FPGA manufacturers do not use metallization for programming. Thus, reverse engineering techniques where attackers identify the silicon mask layer by decapping the chip do not work.

In the case of GPUs (Maurice, C Neumann, & Francillon et al., 2014), Jiang et al. show a correlation timing attack on GPUs to completely recover the AES key. They use 128-bit “electronic codebook” (ECB) mode AES encryption with T-tables, which uses a 16B key to encrypt a 16B block. Each thread performs one block-encryption and the keys are stored in the constant cache. All threads of a warp simultaneously read the same round-key. T-tables are stored in GIM since access to them leads to different memory requests which are serialized and hence, storing them in the constant cache would lead to resource wastage. They record time in two ways: clean measurement, where the warp execution time inside a kernel can be measured, and noisy measurement, where the latency of incoming/outgoing messages can be measured. ECB of different bit lengths are applied and a security function (policy) is formulated. As in the case of Intel AES-NI, beyond improving performance, the new instructions help address recently discovered side channel attacks on AES. AES-NI instructions perform the decryption and encryption completely in hardware without the need for software lookup tables. Therefore, using AES-NI can lower the risk of side-channel attacks as well as greatly improve AES performance.

6. Conclusions

In the highly competitive market of today, the choice of a suitable computing platform poses a significant challenge, as the variables that revolve around it are characterized by great complexity. Hence, metrics such as its performance, energy efficiency, security and ease-of-use carry important weight in such a choice. However, these metrics are often at odds with each other. Evidently, performance-inefficient security solutions may constitute a GPU as an unattractive platform for hardware acceleration of compute-intensive applications. Security issues for FPGAs may have been identified in the past and treated by vendors and researchers, but there have been no common practices established in order for them to be examined (or even treated) until now. The reason for this is that only in the last few years GPUs have been building functional blocks in cloud applications (and not only for edge computing standalone applications). As an increasing number of security vulnerabilities of GPUs come to light, both vendors and users should now carefully weigh the performance advantages of GPUs vis-a-vis their security loopholes. An additional level of security when FPGAs and GPUs are operating in Kernel-based Virtual Machines (Ubuntu-Server) with the CPU/RAM resources specified and deployed in a Kubernetes cluster. Security functions (policies) are formulated when ECB of different bit lengths are applied. Also, it has been shown that AES-NI instructions perform the decryption and encryption completely in hardware without the need for software lookup tables.

Acknowledgement

This work is funded by the Horizon Project AI@EDGE (https://aiatedge.eu/)

References:

Boukhtouta, A., Katabi, D. and Kaafar, H. "Security in Hardware Accelerators for Cloud and Edge Computing".

Ghodke, V. (2021, February 25th). “An Overview — Systems-On-Chips (SOCs) And Their Security Risks”. Security Risks in Systems-On-Chip (SOCs).

INACCEL, Available: https://inaccel.com/cpu-gpu-fpga-or-tpu-which-one-to-choose-for-my-machine-learning-training/

Jiang, H., Fei Y., and Kaeli D., (2016) "A complete key recovery timing attack on a GPU," High Performance Computer Architecture (HPCA), p. 394–405

Joshi, D. and Jain R., "Security and Privacy Considerations for Hardware Accelerators in Cloud and Edge Computing".

Maurice, C., Neumann, O. H. C. and Francillon A., (2014) "Confidentiality issues on a GPU in a virtualized environment," International Conference on Financial Cryptography and Data Security, p. 119–135

NVIDIA, Available: https://www.nvidia.com/en-us/edge-computing/.

Pilato, C., Garg, S. and Wu K., "Securing Hardware Accelerators: a New Challenge for High-Level Synthesis," IEEE EMBEDDED SYSTEMS LETTERS.

Pottie, J. and Kaiser, W. "Security and Privacy in Hardware Accelerators for Cloud and Edge Computing".

Ramakrishnan, K. and Sekar, R. "Securing Hardware Accelerators in Cloud and Edge Computing".

Wang, X., Zhang, Y. and Li X., "Securing Hardware Accelerators in Cloud and Edge Computing".

Hybrid Threat and Information Influence in Connection with Security of Supply

Tehi Palletvuori and Jyri Rajamäki
Laurea University of Applied Sciences, Espoo, Finland

tehi.palletvuori@student.laurea.fi
jyri.rajamaki@laurea.fi

Abstract: Hybrid threat is a multidimensional and hard-to-detect activity. It includes a wide range of actions, from influencing information to the military means by which the hybrid actor achieves its goals. These goals can include weakening or even destroying the target. Security of supply means preparedness and continuity management actions, which aim to safeguard economic activities and related systems that are necessary for the population's livelihood, the country's economic life, and national defense in the event of exceptional conditions and comparable serious disruptions. Both hybrid threat and information influencing can disrupt the realization of the goals of security of supply. This work-in-progress paper proposes a framework, which consists of hybrid threat and its sub-classification, and information influencing as one of the means to implement hybrid threat. The framework also describes the security of supply and elements that are used to combat information influence and maintain the security of supply. In addition, the framework paper discusses what kind of elements measuring the maturity level of an organization's prevention of information influence could consist of.

Keywords: security of supply, hybrid threat, countering information influence, maturity level of information influence prevention, information resilience.

1. Introduction

The ongoing international change in power structures offers a fertile growth environment for hybrid threats. The growing conflict of values between the West and authoritarian states is eroding international norms and institutions and exposing open Western societies to hybrid threats. (Hybrid CoE, 2023) This work-in-progress paper fills the gap that has not been described in the connection between the realization of security of supply goals and hybrid threat and information influencing. The finding of this paper is the proposed framework, which consists of hybrid threat and its sub-classification and information influencing as one of the means to implement hybrid threat. The framework also describes the security of supply and elements that are used to combat information influence and maintain the security of supply. The framework has been created by examining and combining theoretical data. In addition, the framework paper discusses what kind of elements measuring the maturity level of an organization's prevention of information influence could consist of.

2. Framework for hybrid threat and security of supply interaction

According to Iso-Markku (2022) Security of supply is often only seen as guaranteeing the provision of specific materials or goods, such as military equipment or energy supplies. Finnish view on the security of supply is broader. Finnish Security of Supply Act (1992) defines (Figure 1) security of supply as the safeguarding of economic activities and related systems that are necessary for the population's livelihood, the country's economic life and national defense in the event of exceptional conditions and comparable serious disruptions.

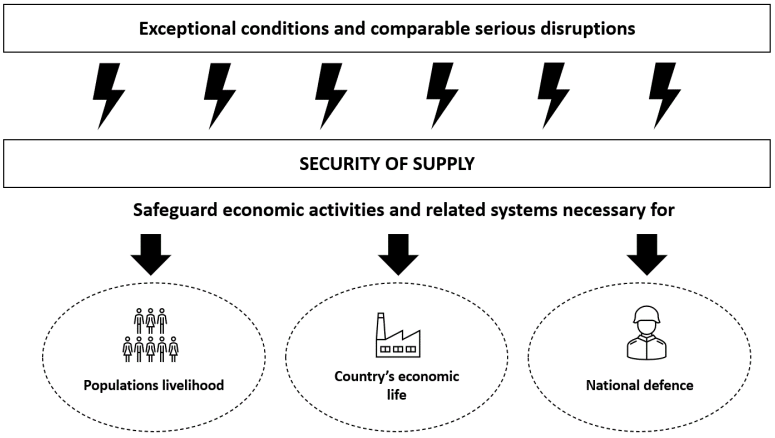


Figure 1: Security of supply (adapted from Finnish Security of Supply Act 1992)

The proposed framework (Figure 2) is divided into five levels. The bottom of the figure is the security of supply, which is the function needed to maintain a functional society during exceptional circumstances and serious disturbances. According to the Finnish Government's decision on the security of supply goals (2018), both hybrid and information influencing can disrupt the realization of the goals of security of supply. By nature, hybrid threats, as well as the threats of the cyber environment, cross borders.

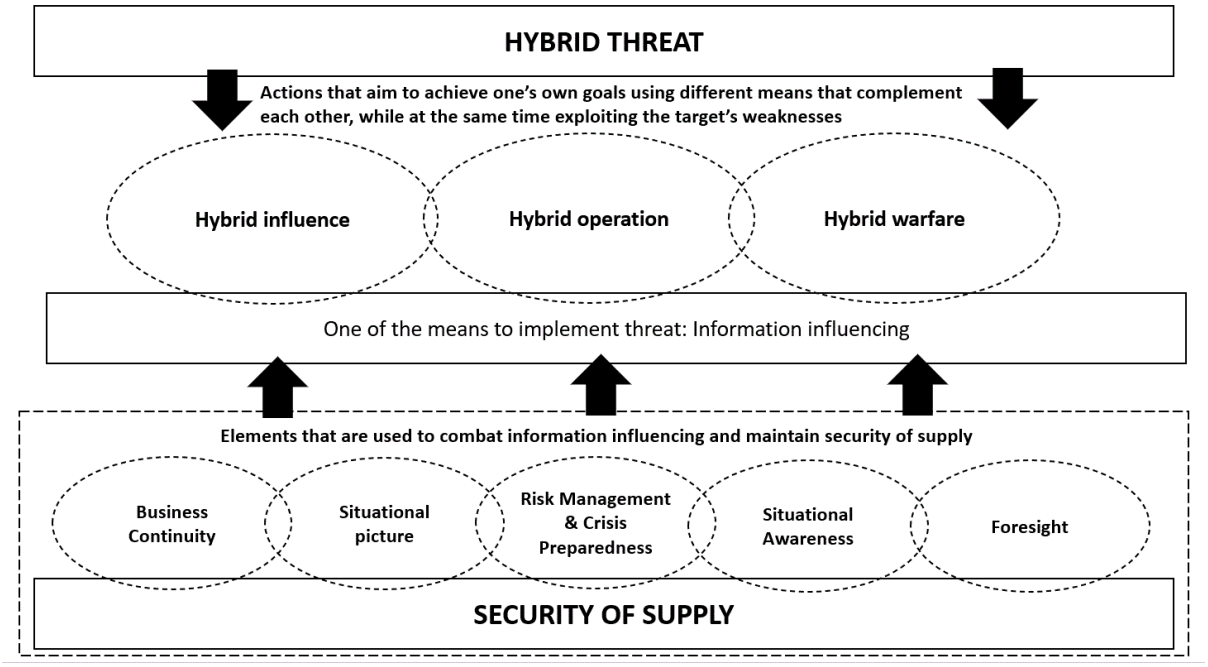


Figure 2: Framework for hybrid threat and security of supply interaction

Information influencing is a systematic activity in which the aim is to implement changes in the target's information and perception environment in accordance with the influencer's own goals through the modification of information (Vocabulary of Comprehensive Security, 2017). This proposed framework focuses on information influencing and elements that are used to combat information influencing and maintaining the security of supply.

2.1 Hybrid threat

On top of the framework (Figure 2) is a hybrid threat and the term relates to conventional or unconventional warfare by which a hostile actor takes advantage of the unclear territory between peace and war (Weissmann, 2021). However, there is no consensus on the definition of the term (Underwood et al., 2022). The problem is also that the terms hybrid threat and/or hybrid warfare are used comprehensively to cover different situations (Weissmann, 2021). Tikanmäki and Ruoslahti (2022) have studied how hybrid terms are discussed in recent scholarly literature. Regardless of the difficulty of defining the term hybrid threat, different actors in society must be able to defend themselves against various hybrid actions. Similar countermeasures may be needed regardless of the name of the term in question. In attribution, it may be more important to specify in theoretical terms what it is about, but in control measures, the functionality and effectiveness of the control measures are essential.

Kalniete and Pildegovičs (2021) interpret the hybrid threat as an umbrella term covering a range of destabilizing and synchronized civil and military actions. These activities may include disinformation and cyber-attacks, causing political or financial corruption, infiltrating agents of influence, pressuring independent media, and purchasing critical infrastructure (Kalniete and Pildegovičs, 2021). Wigell, Mikkola and Juntunen (2021) have proposed hybrid threat classification to be divided into hybrid interference, hybrid operations and hybrid warfare. In this proposed framework hybrid threat is divided accordingly into three parts, which are hybrid influence, hybrid operation and hybrid warfare. Due to the fact, there is no agreement on the definition of terms this proposed framework uses close to self-explanatory terms to divide hybrid threat actions from soft to hard. Influence actions are softer actions than war actions, and operations are in between those two. Term influence

is also used to distinguish this classification from Wigell et al. (2021) classification of which content has not been examined in detail.

2.2 Combat elements

The proposed framework (Figure 2) includes five elements, which are used to combat information influencing. Those elements are business continuity, situational picture, risk management and crisis preparedness, situational awareness, and foresight as the fifth element. Business continuity refers to those measures that enable the organization to manage various disruptions that threaten its operations with pre-planned and implemented arrangements and management tools (Hopkin, 2018). Business continuity management is a strategic and operational preparedness process approved by the top management of the organization that improves the security of supply (Klemm, 2019). According to Hopkin (2018), business continuity planning is a comprehensive management process that identifies potential threats to the organization and the effects on the business that these threats may cause if they materialize. In this model, business continuity management ensures that relevant entities can fulfill their duties.

The vocabulary of Comprehensive Security (2017) defines the situational picture as a compiled description of the prevailing conditions and the events that created the situation at the time it occurred. A situational picture ensures that the position and role of relevant entities are known and thus can be managed accordingly. Situational awareness means an understanding of what is happening or happened, and the circumstances that influence it, as well as the goals of the different parties and possible development options, and it can be based on a situational picture (Vocabulary of Comprehensive Security, 2017). According to Jantunen (2015), disinformation and information encryption, and censorship can be used to influence the situational picture, but influencing situational awareness requires influencing the interpreter of the situational picture. Foresight means the process of developing knowledge and understanding about the future of a specific unit of analysis or actor system (Piirainen and Gonzalez, 2015). In this model situational awareness brings an understanding of the dynamics of the situational picture. Foresight provides the capability to predict and analyze the potential development of the situation.

Risk means the impact of uncertainty on goals (Finnish Standards Association, 2018). Risk management is a systematic activity and includes, in addition to risk analysis, the planning, implementation and monitoring of the necessary measures, as well as corrective measures (Vocabulary of Comprehensive Security, 2017). Measures to strengthen crisis resilience are preparedness for threats related to the well-being and safety of society. These threats include, for example, hybrid threats, the effects of climate change, disasters caused by humans and nature, and the aftereffects of epidemics and pandemics. The joint preparation, planning, training and implementation of various actors in society is carried out in broad cooperation. One important part of crisis resilience is securing the security of supply in all conditions. (Finnish Government, 2020) The four abovementioned combat elements are used for risk management and crisis preparedness, which are the core functions to combat information influencing to maintain the security of supply.

3. Discussion

An individual, authority, private sector actor or non-governmental organization that is subject to hybrid threat is not necessarily the actual target, but rather a tool to achieve the hybrid actor's goal. The proposed framework has focused on information influencing actions on an organizational level, not on an individual citizen level. An organization's prevention of information influence maturity level can be measured. By measuring the maturity level organization can enhance its combat capability against information influencing.

Information influencing maturity level measure elements could be related to defining the roles and responsibilities and information resilience training and countermeasure process. Also, risk management, business continuity management, preparedness, crisis management and situational picture and awareness-related elements could be added to the maturity level self-assessment tool.

The framework will be used to create the structure and different parts of the maturity level assessment tool. The structure and sub-areas of the maturity level of countermeasures against information influence create a basis for further work, where a battery of questions can be created for the measure, which authorities and companies can use when examining the maturity level of their own organization's countermeasure against information influence.

Regarding countermeasures, it is necessary to understand our role as part of the EU and the fact that hybrid threats and their various means know no boundaries, just like cybercrime. The measures to combat hybrid threats are implemented locally in cooperation with authorities and the private sector, as well as non-governmental organizations and citizens. Organizations in both the public and private sectors and non-governmental organizations are preparing for information influence and countermeasures.

For further academic research, an interesting topic would be to further clarify combat elements when countering hybrid threats and information influence that threaten the security of supply. There is no previous scientific research on this subject, and this might be because the security of supply is seen material related functionality.

References

Finnish Government (2020) Government Report on Finnish Foreign and Security Policy, [online], Publications on the Finnish Government 2020:30, https://julkaisut.valtioneuvosto.fi/handle/10024/162513.

Finnish Government decision on security of supply goals (2018) Valtioneuvoston päätös huoltovarmuuden tavoitteista 5.12.2018, [online], Finnish Ministry of Justice, Edita Publishing Ltd., https://www.finlex.fi/fi/laki/alkup/2018/20181048.

Finnish Standards Association (2018) SFS-ISO 31000:2018 Risk Management – Guidelines, [online], 2nd edition, SFS Online, https://online.sfs.fi/.

Hopkin, P. (2018) Fundamentals of risk management: Understanding, evaluating and implementing effective risk management, Fifth edition, Kogan Page Ltd, London.

Hybrid CoE (2023) Hybrid threats as a concept, [online], The European Centre of Excellence for Countering Hybrid Threats, https://www.hybridcoe.fi/hybrid-threats-as-a-phenomenon/.

Iso-Markku T. (2022) The EU and Finland's Security of Supply: A "turn" in EU thinking provides new opportunities, but significant differences remain, [online], Finnish Institute of International Affairs, FIIA Briefing Paper 330, https://www.fiia.fi/julkaisu/the-eu-and-finlands-security-of-supply.

Jantunen S. (2015) Infosota: "iskut kohdistuvat kansalaisten tajuntaan", 2nd edition, Kustannusosakeyhtiö Otava, Helsinki.

Kalniete, S. and Pildegovičs, T. (2021) Strengthening the EU's resilience to hybrid threats, European view, 20(1), 23-33. doi:10.1177/17816858211004648.

Klemm, K. (2019) Huoltovarmuus: Varautumisella Selviytymiskykyä, Tietosanoma, Helsinki.

Piirainen, K.A. and Gonzalez, R.A. (2015) Theory of and within foresight — "What does a theory of foresight even mean?", Technological forecasting & social change, 96, pp. 191-201, doi:10.1016/j.techfore.2015.03.003.

Security of Supply Act (1992) Laki huoltovarmuuden turvaamisesta 18.12.1992/1390, [online], Finnish Ministry of Justice, Edita Publishing Ltd., https://www.finlex.fi/fi/laki/ajantasa/1992/19921390.

Security Strategy for Society (2017) Yhteiskunnan turvallisuusstrategia, Government resolution / 2.11.2017, [online], The Security Committee https://turvallisuuskomitea.fi/wp-content/uploads/2018/04/YTS_2017_english.pdf.

Tikanmäki, I. and Ruoslahti, H. (2022) How are Hybrid Terms Discussed in the Recent Scholarly Literature?, Proceedings of the 21st European Conference on Cyber Warfare and Security, Thaddeus, E., Nabeel, K. and Cyril, O. (Eds.), doi: 10.34190/eccws.21.1.457.

Underwood, A., Emery, A., Haynsworth, P. and Barnes, J. (2022) All Quiet on the Eastern Front: NATO Civil-Military Deterrence of Russian Hybrid Warfare. Joint Force Quarterly : JFQ, 105, pp. 75-85.

Vocabulary of Comprehensive Security (2017) Kokonaisturvallisuuden sanasto TSK 50, [online], Sanastokeskus TSK, The Security Committee https://turvallisuuskomitea.fi/viestinta/kokonaisturvallisuuden-sanasto/.

Wigell, M., Mikkola, H. and Juntunen, T. (2021) Best Practices in the whole-of-society approach in countering hybrid threats, [online], European Parliament, Directorate-General for External Policies of the Union Brussels https://data.europa.eu/doi/10.2861/379.

Weissmann, M. (2021) Conceptualizing and countering hybrid threats and hybrid warfare: The role of the miliary in the grey zone, Hybrid Warfare: Security and Asymmetric Conflict in International Relations, Weissmann M., Nilsson, N., Palmertz, B. and Thunholm, P. (Eds.), doi:10.5040/9781788317795.

Demand Analysis of the Cybersecurity Knowledge Areas and Skills for Nurses: Preliminary Findings

Jyri Rajamäki [1], Paresh Rathod [1] and Kitty Kioskli [2]
[1] Laurea University of Applied Sciences, Espoo, Finland
[2] trustilio B.V, Amsterdam, The Netherlands

jyri.rajamaki@laurea.fi
paresh.rathod@laurea.fi
kitty.kioskli@trustilio.com

Abstract: The purpose of this paper is to present a preliminary analysis of the cybersecurity market demand in the nursing and health sector. Currently, the market demand study is ongoing under the Digital Europe Programme CyberSecPro project, which strengthens the role of higher education institutions as a provider of practical and working life skills. The project promotes reliable digital transformation in critical sectors, such as healthcare. The rapid development of e-health emphasizes the central position of cybersecurity in healthcare organizations that are increasingly the targets of cyber-attacks. This descriptive literature review explores what a nurse needs to know about cybersecurity. Our results show that awareness of cyber risks is weak in the healthcare sector. Understanding cyber risks and recognizing the effects of one's own activities increases the cybersecurity of the entire organization, therefore cybersecurity training for nurses should be increased. Our study suggested that nurses' most important cyber skills are their own cyber-safe way of operations, identifying cyber threats related to equipment, identifying the effects of cyber disruptions, and acting in a cyber disruption situation. Future nurse training programs should be updated to include these skills. Additionally, the teaching of nurses must be developed so that it meets these competence needs.

Keywords: cybersecurity skills, cybersecurity in healthcare, nursing competence, cybersecurity education, cybersecurity training

1. Introduction

As the healthcare sector continues to offer life-critical services while working to improve treatment and patient care with new technologies, criminals and cyber threat actors look to exploit the vulnerabilities that are associated with these changes. Among the most significant cyber-attacks from over the last 5 years are Anthem Blue Cross (78.8 Million Affected), American Medical Collection Agency Data breach (25 Million Patients Affected), Premera Blue Cross (11 Million Affected), Excellus BlueCross Blue Shield (10 Million Affected), and University of California, Los Angeles Health (4.5 Million Affected) (Stoianov & Bozhilova, 2019). So, it is crucial for all healthcare professionals to know how to manage their work in a cyber-secure manner and understand the effects cyber-attacks can have on the healthcare system.

EU Higher Education Institutions (HEIs) have more than 128 cybersecurity academic programs (undergraduate and graduate) as identified by ENISA (CYBERHEAD), JRC (ATLAS), and a variety of reports by the 4 pilot projects (Sparta, CyberSec4Europe, ECHO, CONCORDIA). The academic programs must offer dynamic capabilities and emerging skills that are required in the market to meet the demands for the cybersecurity workforce and expertise. The digital transformation imposes the HEIs to enhance their role in preparing the new generation workforce and to upskill-reskill the existing one in meeting the challenging and ever-growing cybersecurity challenges. Seventeen HEIs and thirteen security companies from sixteen European countries launched the agile CyberSecPro professional cybersecurity practical and hands-on training program that will complement, support, and advance the existing academic programs by linking innovation, research, industry, academia, and SME support. CyberSecPro aims to bridge the gap between degrees, working life, and marketable cybersecurity skill sets necessary in digitalization efforts and become the best practice for all cybersecurity training programs (see Figure 1).

This paper aims to present a preliminary analysis of the market demand for cybersecurity in the nursing and health sector. The ongoing market demand study is part of the CyberSecPro project under the Digital Europe Programme, which focuses on empowering higher education institutions to provide practical and relevant skills for the workforce. This paper also highlights the relevance of the CyberSecPro project in the context of the healthcare sector. In the future, this paper will provide more in-depth insights and report specific outcomes from the ongoing research and innovation work being conducted as part of the CyberSecPro Digital Europe Programme project. Further, CyberSecPro aims to pilot the European Cybersecurity Skills Framework in practice for cybersecurity professional training.

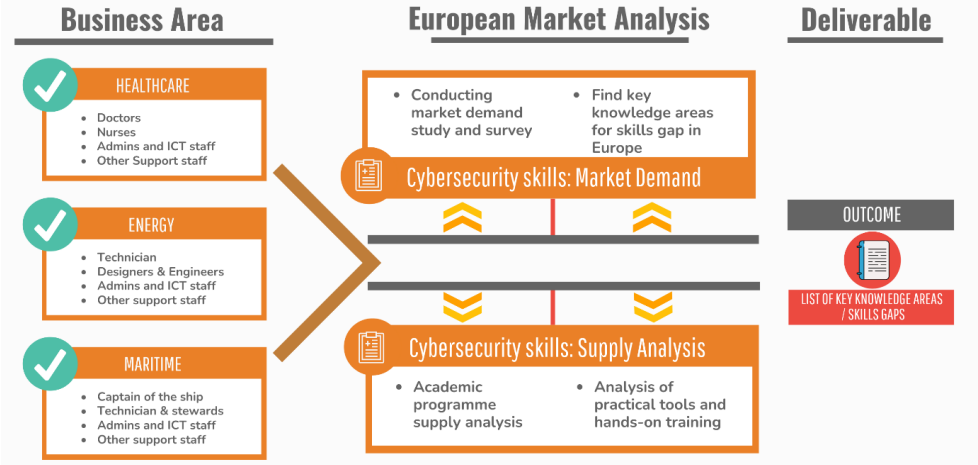


Figure 1: Market analysis of cybersecurity knowledge areas and skills [Source: Rathod & Polemi (2022)]

2. Methodology and collection of research material

The descriptive literature review (Paré et al., 2015) summarizes the most recent empirical research and answers the research questions about *what nurses should know about cybersecurity* and *what cyber skills/competencies they should have*. This work-in-progress paper used Google Scholar as the database. Four different search expressions formed with the Boolean operator were used from the selected search words. Only articles and research reports published in 2022 were considered. Table 1 shows the search results and how many of these are openly available in full text.

Table 1: Search expressions and results

Search expression	Search result documents	Open availability
nurse AND competence AND “cybersecurity in health care”	2	2
nurse AND competence AND “cybersecurity in healthcare”	6	3
nurse AND competence AND “cybersecurity in health care”	5	1
nurse AND competence AND “cybersecurity in healthcare”	52	20

We read the abstracts of 26 articles/research reports that met the search criteria and assessed whether they could find answers to the research questions. Table 2 shows six documents that, based on their abstracts, were selected for closer examination and analyzed using qualitative content analysis (Elo, et al., 2014).

Table 2: Selected articles

Articles/research report	Type	Subject area
Altamimi (2022)	Doctoral dissertation	Medical practitioners’ behavioral justifications
Gioulekas, et al. (2022)	Journal article	Personnel’s cybersecurity awareness
Isännäinen & Tulkki (2022)	Bachelor thesis (In Finnish)	Nurses’ know-how and skills
Rizzoni, et al. L. (2022)	Journal article	Personnel’s cybersecurity training
Sütterlin, et al. (2022)	Book chapter	Human factors
Wasserman & Wasserman (2022)	Journal article	Personnel’s cybersecurity training

In addition, the paper presents the preliminary findings of the Market Demand component of the CyberSecPro project, which falls under the larger Digital Europe Programme. This project aims to enhance the position of higher education institutions as a source of pragmatic and career-oriented competencies, with a particular emphasis on cybersecurity. These early results shed light on the project's initial stages and demonstrate its potential to positively impact the field of cybersecurity education for the nursing and other critical sectors.

3. Results

This paper presents the preliminary findings of an ongoing survey aimed at assessing the current market demand for cybersecurity skills, competencies, knowledge areas, and job roles. The survey focuses specifically on the nursing and health sector and aims to identify the top 12 most critical areas of cybersecurity knowledge required in this field. Based on the initial results of the survey, the following areas were found to be the most important and relevant for the nursing and health sector. It is noteworthy that Risk Management remains the most important and crucial knowledge area.

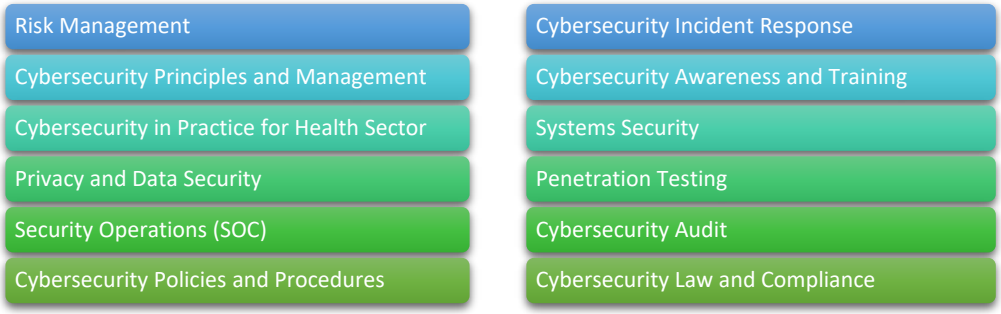


Figure 2: The Most Important Cybersecurity Knowledge Areas for Health and Nursing Sector: CyberSecPro Market Demand Early Outcomes

The further market demand analysis needs more careful data analysis and will be reported in the future long paper that would include the skills, competencies, and job roles. This paper also presents the results of the systematic literature studies.

The results of the qualitative content analysis are reported here according to the 5-step resilience cycle (see Figure 3). In the preparation phase, the nurse's most important skill is to be aware of cyber threats and their effects. In the prevent and protect phases, the most important thing is the cyber-safe operation of the nurse. In the response and recovery phases, the nurse must know how to act in a cyber disruption situation.

3.1 Analyzing and Identifying - Cyber threats and impacts

According to Gioulekas, et al. (2022), only 22.7% of the non-ICT personnel (i.e., doctors, nurses, auxiliary, laboratory, and administrative personnel) felt sufficiently trained in security, while only 38.5% were confident they could recognize a security issue or incident if they encountered one. Identifying the effects of cyber disturbances consists of the effects of cyber disturbances on the patient, the effects of cyber disturbances on the nurse's work, the effects of cyber disturbances on the operation of the hospital, and the effects of cyber disturbances on the medical device (Isännäinen & Tulkki, 2022). As a result of disturbances, the treatment of patients may be compromised in many ways. The likelihood of medication errors increases if prescription systems do not work, or a password is shared illegally (Altamimi, 2022). Stolen personal data are good and long-term investments for cybercriminals, which are used to acquire medicines and devices that are sold on the black market. Personal data is also used to make fake insurance claims. Cyber disruptions are not only a threat to patients' identities, but they can delay or disrupt hospital operations, putting patients' health and well-being at risk. The effects of cyber disruptions directly on the health and safety of patients and the functionality of medical devices. The WannaCry and NotPetya ransomware serve as examples of the deterioration of healthcare performance, as they hindered the operation of implantable heart devices, among other things. The probability of treatment errors increases when the hospital's information and equipment systems are damaged through cyber-attacks. Cyber-attacks are life-threatening crimes because the entire operation of the hospital must be stopped as a result of a cyber-attack at worst, being a big patient safety risk. Cyber-attacks can also affect other critical systems for the hospital, such as heating, ventilation, and air conditioning. These affect not only the patient but also the operation of the hospital and thus directly the nurse's ability to work.

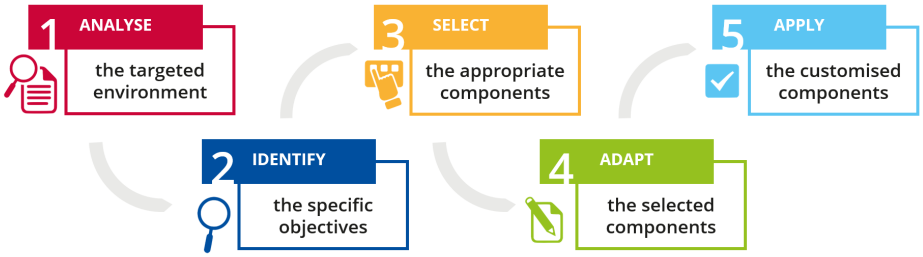


Figure 3: The 5-step resilience cybersecurity application [Source: European Cybersecurity Skills Framework]

Both medical devices and external devices carry risks. The risks of cyber-attacks related to external devices have increased. In the work of a nurse, risks are caused by the use of USBs and personal devices. Both staff and visitors use USB sticks on hospital machines, and anti-virus does not automatically prevent the spread of malware (Isännäinen & Tulkki, 2022). The risk of cyber-attacks increases as hospital equipment increasingly networks with external devices. The staff experiences challenges with devices connected to the Internet. Many nurses do not feel that they are sufficiently prepared or trained to use these devices. However, medical devices may be less at risk than other hospital devices, as hacking medical tools requires basic skills and knowledge of such specialized devices and often an understanding of the clinical implications of device changes (Wasserman & Wasserman, 2022).

3.2 Selecting - The cyber-safe operation of a nurse

Nurses' cyber-safe operation consists of their computer and ICT skills as well as the identification of various risks. The starting point of a nurse's activities with computers is that (s)he knows how to use the ICT operating environment created by ICT professionals in the best possible way.

Identifying the risks associated with usernames and passwords is important. Employees can increase the security of the organization and patients by paying attention to passwords and their use. The workplace password may not be used elsewhere, it may not be written down and it may not be shared with other employees. Double log-in policy, when the same username is logged in on 2 different computers, a security information reminder, and an instruction message should appear to remind the user, and it will require the account holder to log out from the initial session (Altamimi, 2022). The password must not be shared on the phone, which is also partly related to the identification of phishing. The password itself must not contain identifiable information. A strong password created within the limits defined by the organization must be used. Passwords should not be repeated or written down. The use of automatic login and remember me functions on workstations also carries risks. Although sharing passwords may seem like a factor that makes work run smoothly for employees, the unauthorized use of passwords can even be interpreted as criminal activity.

Phishing messages should be recognized. For example, a large Italian hospital with over 6,000 healthcare workers conducts a phishing simulation as part of its annual training and risk assessment (Rizzoni, et al., 2022). Employees can increase the security of the organization and patients by paying attention to suspicious e-mails and their links and attachments. The employee must be considerate when it comes to unexpected emails and learn to recognize suspicious messages. These attachments should not be opened. Random and thoughtless opening of e-mail attachments should be avoided. Recognizing the nurse's own activities and skills, even when under load, is important. Human errors increase in demanding conditions and when deviating from the traditional way of working (Sütterlin, et al., 2022). A stressed nurse is also more likely to open phishing emails. There was a clear positive correlation between nurse workload and opening phishing messages. Healthcare personnel does not know what consequences their behavior may have and what risks it may cause. The staff is not aware that their own actions may contribute to the malware's entry into the hospital's system.

The cybersecurity breaches observed in hospitals are often related to the low cybersecurity awareness of the staff (Gioulekas, et al., 2022). Personnel does not know the consequences of their behavior and they are often only trained further after a cyber-attack has already occurred. Identifying cyber risks is helped by understanding the goal of cyber-attacks. Hackers' main goal is usually not to harm patients but to gain access to the hospital's data network. However, hackers are also interested in patients' individual health and personal information, which usually remains unchanged. These include, for example, blood group, surgeries performed, and social security number. This information ending up in the hands of outsiders always causes a security breach.

Remote work increases the risks of a nurse's work. The studies state that employees were willing to do remote work, even though they are aware of the risks it makes possible. Also, risks associated with using social media applications exist. Although communication channels such as WhatsApp are fast and convenient, they can pose security risks (Altamimi, 2022).

3.3 Adapting and applying - Acting in a cyber disruption situation

A cyber disruption situation may lead to an emergency. In the event of a cyber-disruption, the nurse may not have access to information important to the patient's care. When a cyber-attack is detected, the entire hospital ICT system often must be shut down. The nurse, therefore, must rely on paper-based systems because computer-based systems are down (Isännäinen & Tulkki, 2022).

4. Discussion

Awareness of cyber risks is weak in the healthcare sector (Gioulekas, et al., 2022). Understanding cyber risks and recognizing the effects of one's own activities increases the cybersecurity of the entire organization. For this reason, nurses should be trained more in cyber-safe working. Along with the nurse's cybersecurity activities, the importance of identifying the effects of cyber disturbances emerged from the material. The effects are wide-ranging and affect the entire healthcare organization as well as the patients. Compliance with the security policy improves when medical professionals are involved in the development of cybersecurity practices (Altamimi, 2022). As a follow-up study, the literature research should be expanded into a systematic review over a longer period and from more databases. Further the study will also include more comprehensive market demand analysis. The market demands would be useful to find out the experiences and methods of operation related to cybersecurity from professionals and students in the field. The continuing education needs of nurses already at work should also be clarified.

Acknowledgements

The research conducted in this paper was triggered by the project 'Collaborative, Multi-modal and Agile Professional Cybersecurity Training Program for a Skilled Workforce In the European Digital Single Market and Industries' (CyberSecPro) project. This project has received funding from the European Union's Digital Europe Programme (DEP) programme under grant agreement No 101083594. Special thanks to the partners of these projects and their contributions. The third author (KK) would also like to acknowledge the project 'A Dynamic and Self-Organized Artificial Swarm Intelligence Solution for Security and Privacy Threats in Healthcare ICT Infrastructures' (AI4HEALTHSEC) under grant agreement No 883273. The sole responsibility for the content of this paper lies with the authors. The authors are grateful for the financial support of these projects that have received funding. The views expressed in this paper represent only the views of the authors and not of the European Commission or the partners in the above-mentioned projects.

References

Altamimi, S. (2022). Investigating and mitigating the role of neutralisation techniques on information security policies violation in healthcare organisations. https://theses.gla.ac.uk/82646/1/2022altamimiphd.pdf

CyberSecPro Innovation Project, European Union's DIGITAL-2021-SKILLS-01 Programme under grant agreement no. 101083594 (2022)

Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K., & Kyngäs, H. (2014). Qualitative content analysis: A focus on trustworthiness. SAGE open, 4(1), 2158244014522633. 10 DOI: 10.1177/2158244014522633

European Union Agency for Cybersecurity. (2022). ECSF, European cybersecurity skills framework. Publications Office. https://doi.org/10.2824/859537

Gioulekas, F., Stamatiadis, E., Tzikas, A., Gounaris, K., Georgiadou, A., Michalitsi-Psarrou, A., ... & Ntanos, C. (2022). A Cybersecurity Culture Survey Targeting Healthcare Critical Infrastructures. *Healthcare* Vol. 10, No. 2, p. 327 https://www.mdpi.com/2227-9032/10/2/327/pdf

Isännäinen, A., & Tulkki, S. (2022). Kyberturvallisuus terveydenhuollossa: mitä sairaanhoitajan tulee tietää ja osata? [Cyber security in health care. What should a nurse know?] https://www.theseus.fi/bitstream/handle/10024/755584/Isannainen_Tulkki.pdf?sequence=2

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. Information & Management, 52(2), 183–199. https://doi.org/10.1016/j.im.2014.08.008

Rizzoni, F., Magalini, S., Casaroli, A., Mari, P., Dixon, M., & Coventry, L. (2022). Phishing simulation exercise in a large hospital: A case study. *Digital Health* 8, 20552076221081716 https://journals.sagepub.com/doi/pdf/10.1177/20552076221081716

Scharte, B., Hiller, D., Leismann, T. & Thoma, K. (2014). Summary. In: *Thoma K (ed) Resilien Tech. Resilience by Design: a strategy for the technology issues of the future (acatech STUDY*). Herbert Utz Verlag, München, pp 117–125

Stoianov, N. & Bozhilova, M. (2019). D2.1 Sector scenarios and use case analysis, ECHO, 31 October 2019.

Sütterlin, S., Knox, B. J., Maennel, K., Canham, M., & Lugo, R. G. (2022). On the Relationship between Health Sectors' Digitalization and Sustainable Health Goals: A Cyber Security Perspective. In: *Good Health and Well-Being*, 133 https://library.oapen.org/bitstream/handle/20.500.12657/59127/1/%5BHBK%5D%20Final%20Binder%20196%20pgs.pdf#page=151

Wasserman, L., & Wasserman, Y. (2022). Hospital cybersecurity risks and gaps: Review (for the non-cyber professional). *Digital Health*, 135 https://www.frontiersin.org/articles/10.3389/fdgth.2022.862221/pdf

Hidden Permissions on Android: A Permission-Based Android Mobile Privacy Risk Model

Saliha Yilmaz and Mastaneh Davis
Kingston University, London, UK

k1939123@kingston.ac.uk
mast@kingston.ac.uk

Abstract: The continuously increasing amount of data input on mobile devices has made collating and monitoring users' data not only uniquely personalised but easier than ever. Along with that, mobile security threats have overtaken with rising numbers in bank fraud and personal information leaks. This suggests that there is a significant lack of awareness of security issues among mobile users. Specifically, permission-based passive content leaks are getting more attention due to the emerging issues in data privacy. One reason for this is that permissions are running in the background collecting and transmitting data between applications within the same permission group, without the user's knowledge. This means, that a supposedly innocent application like the Clock, which is linked with the Calendar to provide the date and time functionality, can have access to any other application within the same Calendar permission group, which is compromising confidentiality. Moreover, this can lead to a violation of data privacy as the user is not aware of which assets are being shared between permissions. Developers of mobile platforms have implemented permission-based models to counteract these issues, however, application designers have shown that they are not necessarily complying with the General Data Protection Regulations (GDPR). For the mobile user, this means that app developers, app providers, and third parties who are included in the applications, can gain access to sensitive data without user consent or awareness. To address this issue, this study examines permissions that are inherent in the Android mobile infrastructure and exemplifies how they can reveal delicate user information, identify user behaviour, and can be shared among other applications - without obviously breaching GDPR guidelines. 10 first-party Android applications were statically analysed by their permissions and manually investigated for their actual purpose and privacy risk. Finally, considering the affected area, these permissions were categorised into four asset groups that form the base of a risk model. With risk levels from low to high, this model provides detection of risks on data privacy in mobile permissions and highlights the difficulty with GDPR compliance, which we therefore named PRAM, a permission-based Android Mobile Privacy Risk Assessment Model.

Keywords: Android Permissions, Mobile Security, Data Privacy, GDPR, Privacy Risk Assessment, Data Intelligence

1. Introduction and Background

In the last decade, mobile devices have become increasingly intertwined with every aspect of our life due to the advancement of smart mobile devices providing functionalities similar to personal computers. As a result, the current mobile devices store and transfer more personal data such as location, personally identifiable information (PII), photos, etc. (Onik, 2018). In particular, the Android platform facilitates first- and third-party applications with application programming interfaces (API) that allow the applications to access information and resources of the system and device hardware (Almomani, 2020). This resulted in increasing the risk of leakage of users' personal information (Kim & Oh & Choi, 2023) and mobile malware attacks (Ashawa & Morris, 2020). Research has shown that the openness of the Android system and the popularity of Android mobile devices have driven the fast development of Android applications, whilst this also led to an increase in personal data leakage (Guaman et al, 2021). One of the main reasons for this is the use of "permissionware" – the end-user knowingly downloads mobile applications and accepts the permissions policies but generally finds it difficult to understand whether the apps access certain information in the background (Guaman et al, 2021). The permission system is fundamental to the operation of mobile apps with the primary purpose to control access to sensitive information and protect system resources. The permission selection relies on the users' ability to decide based on the application's functionality. However, in most cases, users are unaware that some applications may give additional permissions to unrelated parties (Alenezi. M & Almomani. I, 2017). In a recent study Alenezi and Almomani (Alenezi. M & Almomani. I, 2017), conclude that 80.3% of Android apps request more permissions than they need. As the permission system control accesses the sensitive information it creates an environment for attackers to perform malicious activities such as accessing users' confidential information. Furthermore, as part of an Android system, many of the third-party libraries are required access to the Android device information to obtain private data such as geographic information, network connectivity, and MAC address (Zhan et al, 2021). Hence, this may expose users' sensitive information to privacy risks without the users' knowledge or App developers' intentions.

An effective approach to protect mobile data privacy is to monitor, classify the permission behaviour, and assess the privacy risk. However, this process can be overlooked by users as the application permission was introduced for the application's accessibility to device information. In addition, some of the mobile permissions collect

sensitive information without fully describing their scope (e.g., custom permissions). According to the official Android site, Android permissions are classified into three protection levels: normal, dangerous, and signature/signatureorsystem permissions. These permissions are set by developers in the manifest.xml file and due to the evolution of the Android operating system resulting in a more complex permission system, the number of normal and signature permissions has increased (Amlomani & Khayer 2020).

Several studies have analysed the privacy risk posed by Android applications (Hamed & Ayed, 2016), (Kim & Oh & Choi, 2023), (Gamba, 2019). The authors Kim & Oh & Choi (2023) have identified privacy indicators and based on them, introduced a model to calculate a privacy score for an Android application. However, the model requires various criteria to evaluate and quantify the objective information i.e., ability, benevolence, and integrity of personal data, and subjective indicators in the interaction between the service and its users, i.e., experience, reputation, and inclination. In another study by Gamba (2019), the privacy of the pre-installed Android applications has been analysed. The study shows that the Android permission system lacks transparency relating to data privacy enhanced by third parties and developers using custom permissions.

The European Union Agency for Network and Information Security (ENISA) addresses this in their privacy and data protection study regarding mobile applications: the European data protection framework defines 'personal data' as identified or identifiable information of a 'natural person', which is very broad (ENISA, 2018). Specific contextual and legal classification in which the information is being processed is needed for an elaborate analysis (ENISA, 2018). This circumstance makes it difficult to determine if mobile application developers' implementation as well as app data collection are following the GDPR guidelines or not. Ideally, the use of an app and processing its sensitive data for a particular specified purpose needs the explicit consent of the user, according to Article 9 of GDPR (ENISA, 2018). Our investigation however has shown that not all mobile permissions comply with this and already fail with the visibility and option to consent to the user. This is not only affecting the more vulnerable third-party applications but also pre-installed first-party apps, our study finds.

To highlight these privacy issues, we considered only pre-installed first-party applications as most research around privacy risk is concerning third-party applications. The thought behind this is that third-party apps are more prone to security vulnerabilities compared to first-party applications, which leaves the pre-installed apps to be overlooked. Therefore, this study focuses on investigating pre-installed Android application permissions as these apps are mostly installed as default on Android mobile devices. Another pivotal reason for us was that first-party applications are more commonly used compared to third-party applications and that regular users are unaware of the apps' possible privacy risks. An example of this is custom permissions, which are often included in applications. With custom permissions, developers can define their own permissions and share resources and capabilities with other apps (Sarkar et al, 2019). This is not only giving developers a leeway to access the functionalities of other applications but raises a privacy issue as upon our permission analysis, most custom permissions' purpose was unidentifiable. In addition, the pre-installed first-party apps may include third-party libraries which can violate user's privacy (Gamba, 2019).

Current research in this area has shown that there is a lack of privacy risk protection regarding mobile applications. This resonates in several areas of the mobile application infrastructure (Sarkar et al, 2019) which leaves the end-users' data vulnerable. While we are highlighting this angle on privacy issues emerging from mobile permissions and remedy the shortcomings of mobile security tools to identify permission vulnerabilities, we also propose a new approach to improve the efforts of application developers to address privacy issues.

2. Data Collection and Methodology

As all Android Operating System devices come with pre-installed applications, 10 of these apps have been used in this study on a mobile device with Google being the main developer (Table 1). These apps were selected on their 'essentiality' and daily use (e.g., Calendar, Clock, Camera). After downloading the applications and their Android Package files (also known as APK), a pen-testing tool named Mobile Security Framework (MobSF) was used to statically analyse the APKs (MobSF Documentation, 2022). We chose MobSF as it is – so far - the closest solution to an automated static analysis tool with a Graphic User Interface (GUI) which makes the data collection of applications easier and more transparent than other pen-testing tools.

The focus of this analysis is the mobile permissions of each application, which were manually collected and categorized into permission groups with a description of their purpose.

The following phases were conducted to disclose the permission misuse and create the variables for the Privacy Risk Model:

1. Display Information

2. Static Analysis
3. Manual Data Evaluation and Data Cleaning
4. Variable Categorization
5. Risk Modelling

The first phase was to investigate what the Android applications from a mobile device with an Android Operating System (e.g., tablet) displayed in the permission settings of the application user interface (see Table 1). The displayed permissions were then compared with the declared permissions in the Android manifest file of each application. For this comparison, the static analysis function of MobSF was utilized. The reason behind this comparison is to see whether the displayed permissions match with what is declared in the manifest file, and if not, to reveal any discrepancies.

We collected with phase A & B data from (so far) 10 applications, cleansed the data from duplicates, and evaluated the findings manually in phase C by investigating the purpose of each of the 351 permissions as these were not clearly described by the Android Developers Documentation. While we included a description of the purpose for each permission, we moved on to phase D and categorised our permission data into 8 variables. It is worth mentioning that our initial categorization of the permission display included 'functional permissions' (permissions, which are necessary to the functionality of certain application activities and somewhat visible to the user) making it a ternary variable but was later changed to the current binary for convenience.

Lastly, we assessed all the collected data and created categories to form a risk model for a permission-based risk assessment proposal in phase E.

3. Permission Analysis

Our initial proposal of the Permission-based Risk Assessment consists of five categories that we declared as the input variables for the privacy risk assessment. These are *Application, Displayed/Not Displayed, Permission Group, Android/MobSF category and Accessed Area.* The *Application* category contains the name of each application, the *Displayed/Not Displayed* are is a binary function (1 for displayed, 0 for not displayed, see figure 3 for an example of permission display), the *Permission Group* (created by us) is showing the affected features of the mobile device, the *Android/MobSF category* reflects the by default declared Android permission protection level and the last category – *Accessed Area* – was introduced by us to determine which data the permission is accessing.

Our initial results show that, after comparing the manifest file declared permissions of each application with the actual display in the permission settings of the mobile device, most permissions are not being displayed or made aware to the mobile user (Figure 1). Moreover, permissions, which are categorized as *dangerous* by Android constitute at least half of the total amount of permissions per analysed app (Figure 2). For a better understanding regarding the Android permission categories: currently Android is classifying its permissions as either *normal, dangerous,* or *signature/signatureorsystem* and denotes these as the *Android Protection Levels* (Android Developers Page, 2023). Normal declared permissions allow access to data and actions that present very little risk to the user's privacy while dangerous permissions give an app additional access to restricted data and allow apps to perform restricted actions that more substantially affect the system and other apps (Android Developers Page, 2023). Signature and SignatureOrSystem permissions can be used between two or more applications that have to share the same declared certificate to ensure privileged access. The reason behind this is that third-party app developers are not granted access to sensitive data. So, if app B wants to access/communicate with the signature permission included in app A, it must share the same signing key (authentication key). However, if two or more applications use the same signature permission these permissions are granted automatically by the system upon installation – without the user's consent (Android Developers Page, 2023).

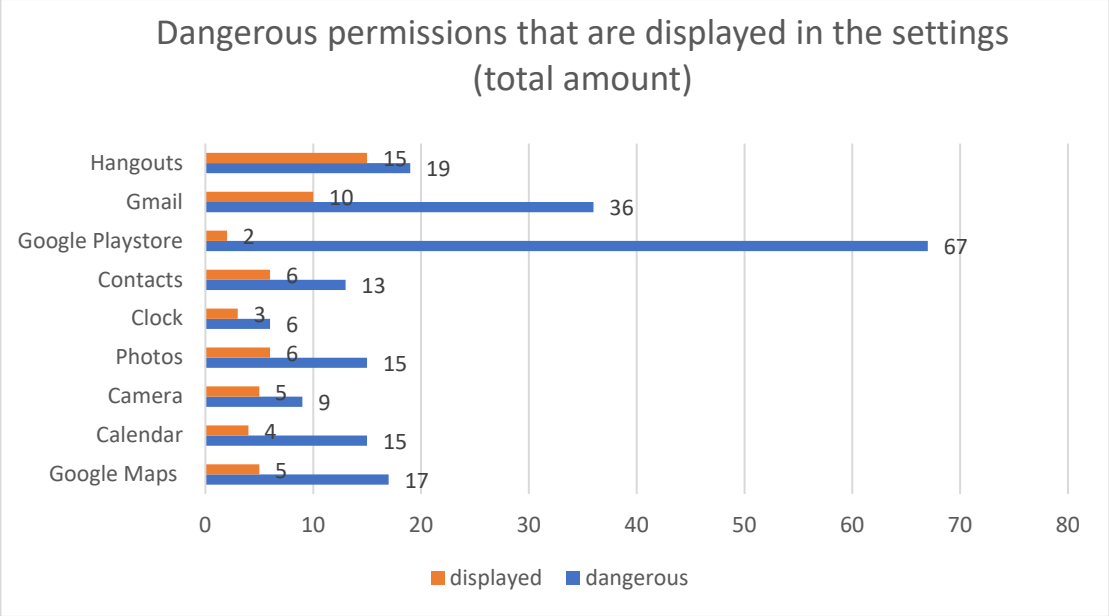


Figure 1: Dangerous permissions displayed in the settings of investigated applications.

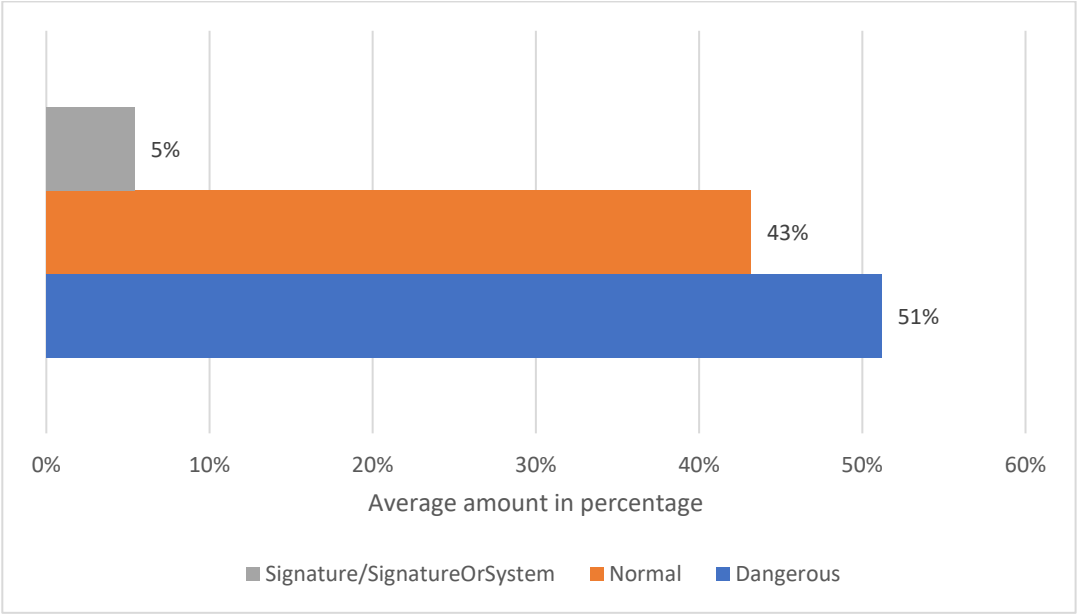


Figure 2: Average amount (%) of signature, normal and dangerous permissions in investigated applications.

When combining these two results, the outcome is that despite having the majority of permissions declared as dangerous (51% of all apps), most are not displayed (72% of all dangerous declared permissions) for the end-user in the mobile device settings. Google Maps for example has only 5 out of 12 dangerous permissions displayed (Table 1).

Table 1: Dangerous and displayed/not displayed permissions of the investigated permissions.

App	dangerous	displayed	not displayed	dangerous and not displayed in %	overall dangerous and not displayed
Google Maps	17	5	12	71%	6%
Calendar	15	4	11	73%	6%
Camera	9	5	4	44%	2%
Photos	15	6	9	60%	5%
Clock	6	3	3	50%	2%
Contacts	13	6	7	54%	4%
Google Playstore	67	2	65	97%	33%

App	dangerous	displayed	not displayed	dangerous and not displayed in %	overall dangerous and not displayed
Gmail	36	10	26	72%	13%
Hangouts	19	15	4	21%	2%
sum of all dangerous permissions	197	56	141		72%
		28%	72%		

There are also functional permissions, that are mostly categorized as normal according to the Android Protection Level. Although the use of functional permissions is conscious to the user, it is not really a choice since the functionality of many applications relies on allowing those permissions – receiving and sending emails would not work without network permissions nor would any map application without access to the location permission. As normal permissions are usually considered to present very little risk to the user's privacy, they can still raise privacy issues. Our analysis has shown that functional permissions such as `android.permission.INTERNET` (allows applications to open network sockets), which are considered as 'normal' under the Android Protection Level, can still be exploited by other apps either through permission misuse or malicious intent.

Further findings suggest that permissions not only run within one application but can access other applications' permissions in the background. Our research has revealed that these unauthorised hidden permissions run in the background collecting and transmitting data without the user's knowledge. This can be explained by Android's Intents, which are messages allowing application components to request functionality from other components within the Android system (Android Developers Page, 2023). When this happens, applications can request – by using Intents – a certain app (here: Google Maps) to start searching for a location or request directions from one location to another. In this example, Google Maps opens and starts to display the location specified in the Intent. The access to this resource is supposedly protected, apps requesting Google Maps to search for a specified location need to declare the "android.permission.ACCESS_FINE_LOCATION" permission.

The predicament here is, that the Android OS does not check if an app that is accessing a permission-protected resource through another app has itself requested that permission (Muttik et al, 2015). The developer of the app that is exposing access to the permission-protected resource is responsible for this check, which leaves a lack of control that can be used by apps to get access to sensitive resources without users' authorisation (Muttik et al, 2015). An example of this is presented in Figure 3, in which one of the tested applications demonstrates the connection chain to other applications in the permissions settings. This chain of collusion can be exploited and lead to information leaks and unauthorised access.

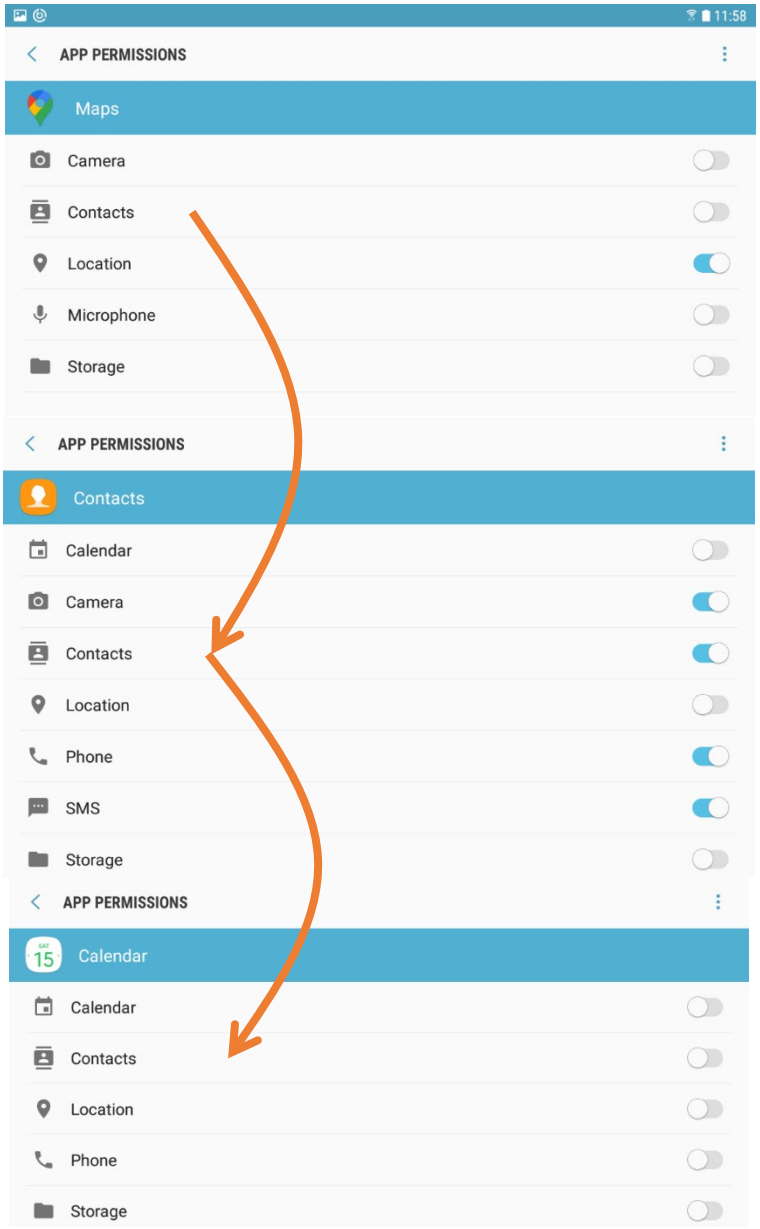


Figure 3: Three colluding permissions that have the potential to access and share data.

4. Permission-based Privacy Risk Assessment proposal (in development)

As mentioned in Chapter 3 the five categories form the basis of our risk assessment model. Depending on the composition of the categories/variables the assessment generates a privacy risk score (Figure 4). So far, we have a scoring level of *low, medium, and high risk*, first for each permission then – depending on the outcome of all permissions within an app – a scoring for the whole application. A low privacy risk score generally indicates that the permissions of an application are declared as normal (present very little risk to the user's privacy according to Android), displayed on the device's settings, and have only access to a digital asset that does not comprise personal information of the user. A high privacy risk score would instead show that the mostly dangerous and/or signature permissions are accessing Personal Identifiable Information (PII) but are not made aware to the user in the device's settings (not displayed). In between these two scores, there is also the medium risk, which is composed of parts from both attributes.

In addition to this, we are currently in the process of including intertwined/colluded applications as a variable, which have the potential to access more within the allowed permission group (as explained in Chapter 3/Figure 3).

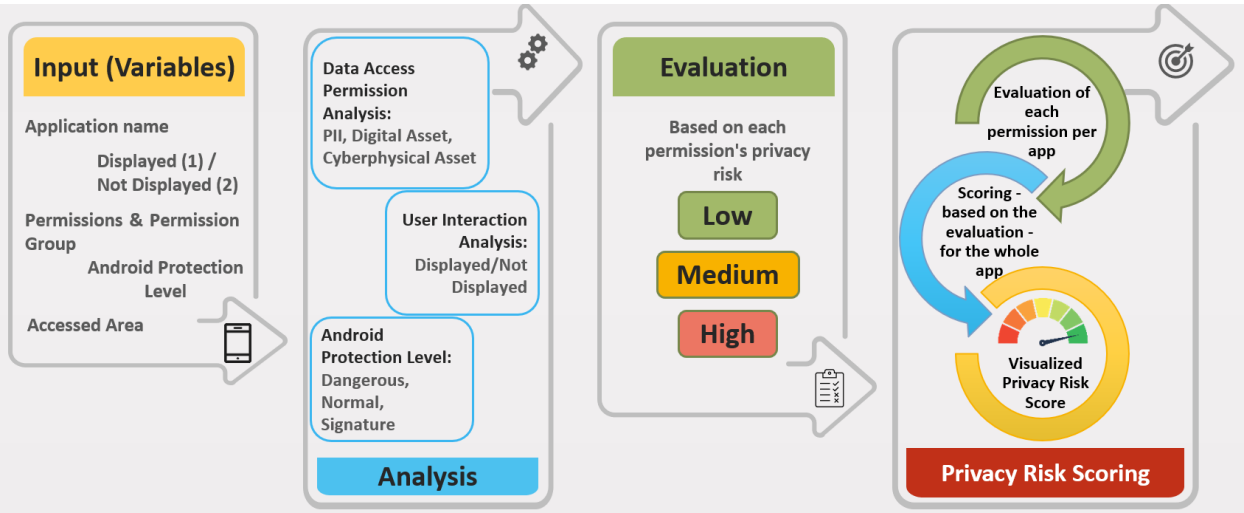


Figure 4: Process of the proposed Privacy Risk Assessment Model.

5. Conclusion

The shortcomings caused by permission models like the least privilege principle violated by coarse-grained permissions (Bugiel et al, 2013), ignorance by end-users due to lack of permission awareness (Zhang et al, 2014), delegation attacks caused by enforcing access control policies at the level of the individual (Ham et al, 2013), and misusing/overusing permissions (Zarni Aung, 2013) have been broadly discussed in several publications. On the one hand, Android users may grant permissions without fully understanding the implications of their decisions and without the ability to balance functionality and privacy. This can lead to unintended outcomes. On the other hand, Android and application developers increase the number of permissions to improve security and ease the access and inter-communication between the applications but turn a blind eye to the flaws these upgrades have inherent. Furthermore, frequently over-privileged permissions are used by application developers without the user's understanding of their exact nature (e.g., custom permissions) with the tendency to increase with each system update.

This study highlights these flaws and investigates the privacy risks of permissions on individual as well as collective app level. So far, this research has shown that there are privacy concerns regarding custom and hidden permissions that have access to delicate information. This goes further with the colliding permissions between applications, in which one declared permission can access more information than initially allowed in the settings. Particularly, it has been shown that some pre-installed applications collect personal and sensitive information without the user's knowledge. Additionally, this study has demonstrated the lack of transparency of data privacy in current policies.

As a result of our investigation, this study introduces a new way of classifying application permissions, based on their purpose. Along with that, this work proposes a risk assessment model based on the aforementioned classification to provide a manually evaluated scoring that informs users and developers about the lack of privacy security and raises awareness of application permissions. Moreover, this risk model will include visualized results by assessing application permissions and informing developers as well as mobile users through a graph algorithm user interface. As an extension of this study, more testing of this model is needed to provide a better solution for the current permission policies on mobile devices. It is anticipated that this study will contribute to the discussion around mobile data legislation and help to address the privacy issues for policymakers.

References

Alenezi, Mamdouh & Almomani, Iman. (2017). Abusing Android permissions: A security perspective. 1-6. 10.1109/AEECT.2017.8257772.

Almomani, Iman & Al Khayer, Aala. (2020). A Comprehensive Analysis of the Android Permissions System. IEEE Access. 8. 216671 - 216688. 10.1109/ACCESS.2020.3041432.

Android Developers. (n.d.). *Android 6.0 Changes*. Available at: https://developer.android.com/about/versions/marshmallow/android-6.0-changes, [Accessed 29 June 2021].

Android Developers. (n.d). *Intent.* Available at: https://developer.android.com/reference/android/content/Intent, [Accessed 29 June 2021].

Android Developers. (n.d.). *Permissions overview*. Available at: https://developer.android.com/guide/topics/permissions/overview, [Accessed 29 June 2021].

Android Developers. (n.d.). *Privacy best practices*. Available at: https://developer.android.com/privacy/best-practices#permissions, [Accessed 13 May 2022].

Ashawa, Moses & Morris, Sarah. (2021). Android Permission Classifier: a deep learning algorithmic framework based on protection and threat levels. Security and Privacy. 4. 10.1002/spy2.164.

Aung, Zarni & Zaw, Win. (2013). Permission-Based Android Malware Detection. International Journal of Scientific and Technology Research. 2. 228-234.

Bugiel, Sven & Heuser, Stephan & Sadeghi, Ahmad-Reza. (2013). Flexible and fine-grained mandatory access control on Android for diverse security and privacy policies. Proc. 22nd Usenix Security Symp.. 131-146.

Xian, Zhan & Liu, Tianming & Fan, Lingling & Li, Li & Chen, Sen & Luo, Xiapu & Liu, Yang. (2021). Research on Third-Party Libraries in Android Apps: A Taxonomy and Systematic Literature Review. IEEE Transactions on Software Engineering. PP. 1-1. 10.1109/TSE.2021.3114381.

European Union Agency for Cybersecurity (ENISA), (2018). *Privacy and data protection in Mobile applications*. Available at: https://www.enisa.europa.eu/publications/privacy-and-data-protection-in-mobile-applications [Accessed 9 May 2023].

Gamba, Julien & Rashed, Mohammed & Razaghpanah, Abbas & Tapiador, Juan & Vallina-Rodriguez, Narseo. (2019). An Analysis of Pre-installed Android Software.

Guaman, Danny & Del Alamo, Jose & Caiza, Julio. (2021). GDPR Compliance Assessment for Cross-Border Personal Data Transfers in Android Apps. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3053130.

Ham, You & Lee, Hyung-Woo & Lim, Jae & Kim, Jeongnyeo. (2013). DroidVulMon -- Android Based Mobile Device Vulnerability Analysis and Monitoring System. 26-31. 10.1109/NGMAST.2013.14.

Hamed, Asma & Kaffel Ben Ayed, Hella. (2016). Privacy risk assessment and users' awareness for mobile apps permissions. 1-8. 10.1109/AICCSA.2016.7945694.

Kim, Nakyoung & Oh, Hyeontaek & Choi, Jun. (2023). A privacy scoring framework: Automation of privacy compliance and risk evaluation with standard indicators. Journal of King Saud University - Computer and Information Sciences. 35. 10.1016/j.jksuci.2022.12.019.

Mobile Security Framework. (2019). MobSF Documentation. Available at: https://mobsf.github.io/docs/, [Accessed 29 June 2021].

Muttik, Igor & Blasco, Jorge & Chen, Tom & Kalutarage, Harsha & Shaikh, Siraj. (2015). Android – Collusion Conspiracy.

Onik, Md Mehedi Hassan & Al-Zaben, Nasr & Yang, Jinhong & Lee, Nam-Yong & Kim, Chul-Soo. (2018). Risk Identification of Personally Identifiable Information from Collective Mobile App Data. 71-76. 10.1109/iCCECOME.2018.8659213.

Sarkar, Anirban & Goyal, Ayush & Hicks, David & Sarkar, Debadrita & Hazra, Saikat. (2019). Android Application Development: A Brief Overview of Android Platforms and Evolution of Security Systems. 73-79. 10.1109/I-SMAC47947.2019.9032440.

Zhang, Yuan & Yang, Min & Xu, Bingquan & Yang, Zhemin & Gu, Guofei & Ning, Peng & Wang, Xiaoyang & Zang, Binyu. (2013). Permission Use Analysis for Vetting Undesirable Behaviors in Android Apps. IEEE Transactions on Information Forensics and Security. 9. 611-622. 10.1145/2508859.2516689.

Late Submission

From Provoking Emotions to fake Images: The Recurring Signs of fake news and Phishing Scams Spreading on Social Media in Hungary, Romania and Slovakia

Kenyeres Attila Zoltán[1] and Lauren Weigand[2]
[1]Eszterházy Károly Catholic University, Eger, Hungary
[2]Duquesne University Kline School of Law

kenyeres.attila.zoltan@uni-eszterhazy.hu
weigandl1@duq.edu

Abstract: The phenomenon of fake news and media manipulation has always existed in human history, long before the invention of digital technology. However, never before in the history of mankind has it been possible to spread fake news so quickly, in such large quantities and to such large masses, as now, in the age of the internet and social media. In this paper we identified 31 recurring signs of fake news and phishing scams spreading on social media in Hungary, Romania and Slovakia, based on the content analysis of 866 screenshots of social media posts, internet articles, phishing emails and SMS messages from these 3 countries. The most common group of signs are signs of provoking emotions. The second largest group of indicators include the characteristics of the media publishing the news. The third major category is the visual appearance and wording of the news. The fourth group of recurring signs refers to the original source of the news. The fifth group of indicators is the lack of reliable and/or official media coverage of the story. The elements of the sixth group of signs are the photoshopped and re-framed 'proof' images and videos that appear in the news. The seventh, and final group, of indicators refers to the prior beliefs and biases of the target audience. Provoking emotions, and thereby turning off the recipient's critical thinking, is the most common sign of fake news, scams and other hoaxes. Consequently, there is a great need for a high level of critical thinking and information literacy regarding social media contents on the part of the recipient. Our research was based on a fake news database collected in the framework of an international Erasmus+ project called "Media Detective". The aim of the project is to develop media literacy training modules for teachers and youth workers that could be used in school settings.

Keywords: fake news, social media, Hungary, Romania, Slovakia

1. Definition of fake news

The phenomenon of manipulation and fake news existed in media long before the invention of digital technology. There are numerous examples in print media. For example, Tengely (2001) analyzed the content of contemporary press reports on the 1913 sword duel between two Hungarian prime ministers (István Tisza and Mihály Károlyi). The researcher analyzed the extent to which newspapers of different political parties distorted reality. Thus, we are really dealing with a historical phenomenon when we discuss fake news and media manipulation. Yet, there is an almost endless debate in contemporary international literature on the definition of what we call nowadays "fake news". The controversy is not unfounded, as the term itself is difficult to grasp and interpret with a single clear definition. Allcott and Gentzkow define fake news as *"news articles that are intentionally and verifiably false, and could mislead readers"* (Allcott and Gentzkow, 2017:213). Veszelszki (2017) adopts this definition for a broader understanding of fake news, but narrows the definition to emphasize the pursuit of financial gain and create the appearance of credible journalism. She includes various internet scams (spamming, phishing, hoax, clickbait), news parody, and propaganda among the concepts related to fake news. She also discusses urban legends as a subtype of fake news. Royster then defines fake news as *"inaccurate articles published online in the guise of a genuine news story that are created without any concern as to their truth or falsity."* (Royster, 2017:276). Klein and Wueller also define fake news as *"the online publication of intentionally or knowingly false statements of fact."* (Klein and Wueller, 2017:6). They do not limit it to articles– fake news can also be video or graphics. Additionally, Pennycook and Rand consider the online publishing platform as one of the most important aspects. They define fake news as *"news content published on the internet that aesthetically resembles actual legitimate mainstream news content, but that is fabricated or extremely inaccurate. Also referred to as false, junk, or fabricated news."* (Pennycook and Rand, 2021:389). Disinformation, misinformation, hyper-aggrandized news and "yellow journalism" are also mentioned as related categories. Researchers Baptista and Gradim wanted to find a clear definition of fake news, so they conducted a content analysis of academic articles published on the topic between 2016 and 2020. They found 41 scientific articles in the Web of Science database and 22 in the Scopus database that attempted to provide a precise definition of fake news. Based on their analysis, fake news is: *"a type of online disinformation (1), with (2) misleading and/or false statements that may or may not be associated with real events, (3) intentionally created to mislead and/or manipulate a public (4) specific or imagined, (5) through the appearance of a news format with an opportunistic*

structure (title, image, content) to attract the reader's attention, in order to obtain more clicks and shares and, therefore, greater advertising revenue and/or ideological gain". (Baptista and Gradim, 2022:640).

Thus, based on our literature review above, we can summarize the concept of current digital "fake news" as follows: 1. fake information that is created on a digital device and published on an online platform; 2. in the form of a news article and/or image and/or video; 3. usually related to well-known people, public figures or controversial current events that are of high interest; 4. the news presents itself as reliable though in actuality it is fake; 5. it appears on dubious websites/social media sites with unverified backgrounds, without a classical editorial team behind it that adheres to ethical journalism rules and without a detailed imprint; 6. the author is intentional, i.e. he/she does not spread the lie by accident or out of ignorance/negligence, but knowingly, most often for political/ideological or commercial gain; 7. the main distribution platform is social media, where users who spread the fake news are not necessarily aware that they are spreading fake news.

2. Phishing and scams

Phishing, in its simplest definition, is data fishing, where fraudsters trick users into providing various personal details in order to abuse them– most often to steal their money. Liu et al. (2011) describe phishing as an ongoing "semantic attack" that tricks victims into sharing sensitive information unintentionally. Saberi et al. (2007) distinguish between "phishing" and "scam". "Phishing" is the process of data theft itself, the tool of which is "scam". They describe phishing as a form of "identity theft", in which fraudsters try to trick the victim into providing confidential information such as online bank account details. A typical process is when the attacker deceives the victim with a spoofed email, which is the "scam". According to Chaudhry et al. phishing *"is a form of cybercrime that aims to deceive users into providing personal and/or financial information, or to send money directly to the attacker. A phishing attack is generally initiated via some form of message, which includes a link to a deceptive domain name which appears to be a legitimate site but is actually controlled by the attacker."* (Chaudhry et al, 2016:247). Wash provides a more general definition of email phishing when he writes: *"a phishing email is an email that pretends to be something that it is not, in order to get its recipient to take an action that they otherwise wouldn't do"* (Wash, 2020:1). Baykara and Gürel also consider phishing to be a form of cyber-crime, where the attacker poses as a real person or institution who *"sends malicious links or attachments through phishing e-mails that can perform various functions, including capturing the login credentials or account information of the victim. These e-mails harm victims because of money loss and identity theft."* (Baykara and Gürel, 2018:1). However, according to Hong (2012) phishing can happen not only via email, but also via SMS, chat, voice messages, multiplayer online games, and various social media platforms. To summarize, in a phishing attack, the attacker(s) collect sensitive customer data (e.g. user account login credentials, credit/debit card numbers, etc.) via spoofed emails or fake websites (Basit et al, 2021).

3. About this research

In our modern digital environment, there is an increasing need to develop critical thinking skills towards media content in order to recognize fake news and other scams. Schools and teachers have a major role to play in developing these skills among students. Nowadays, as Molnár et al (2017) mention, almost all students have smart devices, which can also be used in education for pedagogical purposes, depending on the individual creativity of teachers. Smart devices and other digital equipment can be important tools for developing critical skills for assessing media content and supporting visual learning (Molnár and Szűts, 2015), as scams and fake news often carry visual signs. Digital pedagogy can be an excellent tool to develop a critical attitude towards fake news and phishing scams, since on the one hand, scams are also spread in such digital environments, and on the other hand, as Szűts (2020) writes, digital pedagogy can help students to develop their knowledge in a wide range of ways, building on their creativity, by using digital tools they use in their everyday lives (Szűts, 2020). Digital education is also possible and applicable for adults. Learning circles organized through online tools allow the development of a wide range of skills (Simándi, 2020), including a critical attitude towards media content. There are also opportunities for multi-disciplinary skills development even among the elderly through learning circles (Simándi and Oszlánczi, 2018). Improving media literacy among youth communities is the aim of the "Media Detective" project, which produced a fake news database as the basis for this research. In this paper, we summarize the recurring signs of fake news and other scams spreading on social media by analyzing 866 screenshots of social media posts, internet articles, phishing emails. and SMS messages from Hungary, Romania and Slovakia. This content analysis is based on an international media-literacy project called "Media Detective for Young People: Recognizing and Handling Fake News and Media Manipulation" carried out in Hungary, Romania and Slovakia, funded by the Erasmus+ program in 2022 and 2023. One of the main goals of this project is to develop teaching materials that will enable teachers to show Hungarian, Romanian, and Slovakian students how to recognize fake news and phishing scams.

4. The recurring signs of fake news and phishing

In our content analysis, we were able to classify the signs of fake news and phishing scams into seven main categories, as follows:

4.1 Emotional manipulation: stimulating emotions

Perhaps the most common sign of fake news and phishing scams is that they evoke strong emotions. This is a way of "turning off" users' critical thinking. Montesi (2021) also points to this when he writes that in situations of uncertainty and high emotions, the spread of fake news increases dramatically (Montesi, 2021). According to Albrecht et al. (2010) when news is about us, it affects us on an especially high level, such that we will likely react emotionally. But when it is about "other" [non-personal] things, more physically and/or psychologically distant, we tend to think analytically. The signs of emotion-based manipulation are summarized in the table below:

Table 1: Recurring signs of fake news and phishing scams based on emotional manipulation

If something is too outrageous and/or scares us	For example, an SMS that your credit card has been blocked. Or an outrageous statement a politician never said. Or an alarm about a non-existent threat. Or a phishing email in the name of the national tax office scaring us into opening an attached file (which actually contains a virus or spyware).
When the message sets a deadline and urges us to do something	For example, a phishing SMS sent on behalf of Netflix telling you that you must provide your details within 24 hours to a suspicious link or your account will be suspended.
When the news is "too good to be true"	The fake news that all the members of the Croatian national football team donated the proceeds from the World Cup to charity for children in need has been all over the Hungarian and Romanian press.
If the news is extremely tear-jerking, touching or pleading for help	Social media posts of seriously ill children (mainly cancer patients) pleading for help. You can help sick children by liking and sharing. Such posts are also appearing on social media sites in Hungary, Romania, and Slovakia.
If a prize is too tempting and too easy to get	Offering unrealistically high value prizes (e.g. expensive cars, caravans, furniture, high-value shopping vouchers, iPhones etc.) that can be won for a like and share. In reality, there is never a draw, the aim is to increase the number of followers extremely quickly.

(Source: our own research results)

4.2 The medium that reports the news

The other large group of indicators of fake news and phishing scams includes the characteristics of the media publishing the news. These are the signs of suspicious media outlets:

Table 2: Recurring signs of fake news and phishing scams based on the medium that reports the news

If the link or the social media profile has an unknown and/or frivolous name	We don't often hear about the “NewsHat” website or the “Moneybasket” Hungarian Facebook page. Not by accident, as these are fake news sites. Always be suspicious if the news is not from a well-known, reliable, long-established news outlet.
If the URL does not have a country-specific ending	If the link does not end in ".ro" in Romania, ".hu" in Hungary or “.sk” in Slovakia, it is always suspicious. In the case of websites in Hungary, most of the fake news sites are not ".hu", but ".net", ".eu", ".com", ".me" etc. The reason is that the rules for creating ".hu" sites are stricter, more data must be provided, they are run from Hungarian servers, so they are easier for the Hungarian authorities to check and track down. For sites with other extensions, the rules are less strict, which is why they are so popular with fraudsters.
If the URL reminds you of a familiar/trustful company name, but the ending is unusual	For example: the official website of the supermarket chain Tesco in Hungary is "tesco.hu". The website "tesco-hu.com" is created by fraudsters. The official website of RTL television in Hungary is "rtl.hu", and fraudsters have created a website at "rtl-klub.me".
Copied websites - but with a different URL	A common form of fraud in all three countries is for phishing scammers to copy the official website of a bank or a trusted news site. At a glance, it looks just like the original, official site. However, the URL debunks the scam, because it immediately shows that you are not on the official site.
If there is no “about us” or “contact us” section of the website or it is incomplete	The “about us” section on the websites of reliable news sources contains details of the editorial office: its physical address, telephone number, online contact details, staff names, titles, personal contact details– in many cases even their photo. If there is no "contact us" or "about us" menu item with detailed information, it is a sign of unreliable sources.

Features of the social media profile	For example: If the followers of the social media profile also consist of suspicious/famous sites. Or if the social media profile has only just been created and is already giving away big prizes.

(Source: our own research results)

4.3 The visual appearance and wording of the news

The next major category is the visual appearance and wording of the news. That is, how it is written, what grammatical errors it contains, what design it has. The following table summarizes the external signs of fake news and phishing scams:

Table 3: Recurring signs of fake news and phishing scams based on the visual appearance and wording of the news

Unusual formatting: misuse of small and capital letters, unusual punctuation, too long of a title	In almost all cases, the unusual formatting marks are indicative of fake news. Examples include titles consisting of 3-4-5 lines. A professional journalist would never give a title without a headline. However, these signs are increasingly disappearing.
Spelling mistakes, punctuation errors, incorrect wording	Some of these errors are due to the fact that international scams do not use professional translation software to produce country-specific translations. These signs are also on the decline.
Mocking/sarcastic tone	In the case of heavily partisan, biased news, and fake news, a sarcastic and mocking style is often used.
Words and adjectives referring to spirituality, miracles	For example, the Hungarian Facebook page "Wonderful World", which posted a fake news story about rain in Brazil, which put out forest fires that were raging at the time. Or the Facebook post about the miraculous spiritual journey of bees, which is actually based on a photo stolen from a case in England.
"Amen" or “Done” comments lining up	"Amen" comments are often found under fake posts and fictional touching stories that collect likes, while "Done" comments are often found under fake sweepstakes.
Ask to share/agree/like	A recurring sign of fake news is that it asks to be shared. Professional journalists never ask to be shared, even in cases of missing persons. Be immediately suspicious if you come across a post requesting a share.
Two pictures side by side (in Hungarian cases)	In the case of Hungarian fake news, a recurring feature was the appearance of two (or sometimes three) pictures side by side in the post. This was very rarely seen in Romania and Slovakia.

(Source: our own research results)

4.4 The origin of the information itself

The next group of recurring signs of fake news and phishing scams refers to the original source of the news. Here, we are not referring to the medium or social media profile that shared the news, but to the original source of the information. The following table summarizes suspicious signs about the original source of the news that point to fake news or phishing scams:

Table 4: Recurring signs of fake news and phishing scams based on the origin of the information itself

If the origin is incorrect or missing, or if it just "came from somewhere"	For example: when a friend forwards you an SMS warning you about a rise in petrol prices, but it is unknown where this information originated. Or you receive a Messenger text warning of a weather threat, but you don't know who is the source of this information and the weather service it refers to doesn't actually exist.
If you were selected out of the blue and won something (Instagram, email, SMS)	A common trick used by phishing scammers is to inform you that you have suddenly won something. Often on Instagram, we receive tagging in fake posts announcing a win.
Unusual activities, messages and posts from your friends’ social media profiles	Suddenly, from a friend who never writes, send us a link out of the blue and asking if we are in this video. It's suspicious when one of our friends starts suggesting dream-like bitcoin investment, who has never done anything like this before. Or shares a post with many other friends tagged. It's also possible that our friend's social media page has been hacked and the message is coming from scammers.
Vague reference to authorities without verifiable specifics	A common trick among fake news authors is to refer to well-known organizations, authorities, official bodies, governments (e.g. NASA). However, there is no evidence of it on the official pages of these organizations. If there is a link in the fake post, it is broken or points to an unofficial/fake site.

(Source: our own research results)

4.5 Lack of reliable and/or official media coverage

The next group of recurring signs that indicate fake news are those that indicate sensational and/or shocking news and revealed secrets that appear exclusively on social media and/or unreliable websites. Different media outlets may be reliable for everyone, but generally speaking, if a sensational news story is not reported by a major, well-known, long-established news outlet with professional journalists or official sources, we should be suspicious. The following signs may indicate this phenomenon:

Table 5: Recurring signs of fake news and phishing scams based on the lack of reliable/mainstream and/or official media coverage

Although the news is super-sensational and/or it affects many people, it is not reported by reliable international news sources, and/or official authorities	For example, when a private Facebook profile announces that after 112 years it has snowed again in Cairo and the pyramids are covered in white. If it had, it would have been reported by the entire international mainstream media. Or a fake news story that people born between 1951 and 1989 are receiving substantial discounts from a Romanian bank– but there is nothing about it on the bank's official website. Or a Facebook post about kidnappers in a village, while the police know nothing about it.
If the news promises to reveal secrets, confidential or inside information	When a post or article or message promises to reveal insider, confidential, secret information, you should always suspect that you are being scammed.
If a layman shows you some really sensational/amazing "lifehack" – especially prevalent in videos and on TikTok	One has to be very careful with videos and challenges on TikTok, as they can often cause damage, injury and accidents. Before anyone tries any dangerous chemical or physical processes at home, they should ask a professional. There are also lots of useful videos, many of them harmless, but some of them can be dangerous. Only trust serious, professional, official sources and verifiably credible professionals, or if the layman in the video is referring to serious sources that actually exist.

(Source: our own research results)

4.6 The photos and videos used in the posts: stolen, photoshopped and/or re-framed images

The next group of signs of fake news and hoaxes are 'proof' images and videos that appear in the news. In most cases, these are images and videos stolen from elsewhere, photos and videos that have been edited, or photos and videos that have been re-framed. These signals are summarized in the table below:

Table 6: Recurring signs of fake news and phishing scams based on the photos used

Photoshopped pictures	For example, a photoshopped photo was circulated on international social media showing Hungarian police officers using herds of swine to keep Muslim migrants away from the border. Among Muslims, pigs are considered unclean animals and are therefore shunned. However, the original photo was taken in the Philippines after a typhoon passed.
Re-framed images	Real photos, originally published in a different context. For example, "photo evidence" circulated on Hungarian and Romanian social network sites showing a lot of useless electric cars in Paris because the batteries are too expensive to replace. But the original photo was taken in China, which had nothing to do with expensive batteries.
Stolen photos used for emotional click-bait	Usually emotionally triggering photos of unfortunate and/or elderly people and/or sick or missing children, natural disasters, extraordinary phenomena (e.g.: stolen photos of special works of art). Their publishers call for sympathy, help, sharing and likes. Google Lens shows what other sites have used the image.
Re-framed videos	Videos that were originally made in a completely different place, at a different time, in a different context, possibly from video games. For example, a video allegedly showing Ukrainians pretending to be dead under body bags, as one of the "dead" emerges out from under the bag. The video was originally taken years earlier in Vienna during a demonstration against climate policy.

(Source: our own research results)

4.7 Building on the prior beliefs and biases of the target audience

When a news item tells us exactly what we want to hear, we should immediately become suspicious. This can also be a sign that we are dealing with fake news. Creators of fake news know well what their target audience wants to hear, and they say exactly that. They are exploiting our own preconceived beliefs, political, scientific and other views.

Table 7: Recurring signs of fake news and phishing scams based on the prior beliefs and biases of the target audience

When something suspiciously matches our prior expectations and stereotypes	Even the mainstream media reported the fake news that a Russian tank had been stolen by Ukrainian Roma. Many people have the stereotype that Roma are tricky people. This led many to believe that they had actually stolen a Russian tank. In reality, all we have seen is a re-framed video, originally showing Ukrainians towing a stalled combat vehicle with a tractor. Another fake news story is that Elon Musk spent Halloween in Dracula's castle in Transylvania. The super-rich and eclectic Elon Musk can be imagined deciding to spend Halloween in one of "Count Dracula's" castles. It fits in with the image of him.
When something fits perfectly with our political and other views/biases	For example, a fake Facebook post showing the results of an unspecified "scientific experiment" on how harmful microwaves are. Those who already think microwaves are harmful will believe the news even if every detail of the post tells them that it is about an experiment that never happened.

(Source: our own research results)

5. Summary

We found few differences between the fake news posts and internet scams we analyzed in Hungary, Romania, and Slovakia. Accordingly, it is worth watching for these aforementioned signs of fake news, as they were observed in all three countries. This is no small task. As well, we must be able to step back from our own preconceptions, our own worldview. In relying on them, we can easily fall victim to scams. In total, we were able to identify seven signs that rely on evoking emotions and prior convictions of the receiver. In addition to these, it may be easier to recognize other signs that come either from the source of the news, the medium that published the news, or from the images/videos presented. In this study, we identified a total of 31 signs of fake news and other scams in 7 categories. It is important that a fake news post or phishing scam usually involves several contemporaneous signs. Sometimes, it is not always possible to determine for sure whether news is fake or real based on a single sign. Several factors must always be carefully weighed. A "healthy" level of suspicion is needed, because it is also dangerous not believing in any media. For this reason, it is also very important to know what details to look out for to determine whether the news or the SMS or e-mail you receive is real. Some of these signs are disappearing (e.g. spelling mistakes, bad wording) because, as artificial intelligence develops, scams are becoming more sophisticated and harder to detect. Deepfake videos also fall into this category. This makes it essential to develop media awareness and heighten the critical faculties of the public.

References:

Albrecht, K., Volz, K.G., Sutter, M., Laibson, D.I. and Von Cramon, D.Y. (2010) "What is for me is not for you: brain correlates of intertemporal choice for self and other",[online], *Social Cognitive and Affective Neuro Science*. Vol. 6. No. 2. pp. 218–225. https://scholar.harvard.edu/sites/scholar.harvard.edu/files/laibson/files/what_is_for_me_is_not_for_you_brain_correlates_of_intertemporal_choice_for_self_and_other.pdf

Baykara, M. and Gürel, Z.Z. (2018) "Detection of phishing attacks", [online], *6th International Symposium on Digital Forensic and Security (ISDFS)*, Antalya, Turkey, pp. 1-5. https://ieeexplore.ieee.org/abstract/document/8355389

Baptista, J.P. and Gradim, A.G. (2022) *A Working Definition of Fake News*. [online], https://www.mdpi.com/2673-8392/2/1/43

Basit, A., Zafar, M., Liu, X., Javed, A.R., Jalil, Z. and Kifayat, K. (2021) "A comprehensive survey of AI-enabled phishing attacks detection techniques" [online], *Telecommunication Systems* Vol. 76. 2021. pp. 139–154. https://link.springer.com/article/10.1007/s11235-020-00733-2

Chaudhry, J., Chaudhry, S.A. and Rittenhouse, R.G. (2016) "Phishing Attacks and Defenses", [online], *International Journal of Security and Its Applications* Vol. 10, No. 1 pp.247-256. http://dx.doi.org/10.14257/ijsia.2016.10.1.23

Hong, J. (2012) "The state of phishing attacks", [online], *Communications of the ACM* Vol. 55. No. 101. pp 74–81. https://doi.org/10.1145/2063176.2063197

Klein, D.O. and Wueller, J.R. (2017) "Fake News: A Legal Perspective", [online], *Journal of Internet Law.* Vol.20.No.10. pp.5-13. https://deliverypdf.ssrn.com/delivery.php?ID=348094069068099084119082124113004024022087061054024018026093004006065028028031105106002063005047108007015085105079086070118021046053082082007103072086071102015122030095085017074113113089083084111069105078096121014117095085004102088103081088013029101119&EXT=pdf&INDEX=TRUE

Molnár, Gy. and Szűts Z. (2015) "Visual Learning - Picture and Memory in Virtual Worlds" In: Benedek, A. and Nyíri, K. (eds) *Beyond Words : Pictures, Parables, Paradoxes* Frankfurt, Germany : Peter Lang Verlag. pp. 153-161.

Molnár, Gy., Sik, T.D., and Szűts, Z. (2017) "IKT alapú mobilkommunikációs eszközök és alkalmazások módszertani lehetőségei a felsőoktatásban" In: Mrázik, J. (ed) *A tanulás új útjai*, Budapest, Hungary, Magyar Nevelés- és Oktatáskutatók Egyesülete (HERA) pp. 285-297.

Montesi, M. (2021) "Understanding fake news during the Covid-19 health crisis from the perspective of information behaviour: The case of Spain", [online], *Journal of Librarianship and Information Science* Vol. 53. No.3. pp. 454–465 https://journals.sagepub.com/doi/pdf/10.1177/0961000620949653

Liu, G., Xiang, G., Pendleton, B.A., Hong, J.I. and Liu, W. (2011) "Smartening the Crowds: Computational Techniques for Improving Human Verification to Fight Phishing Scams". [online], Symposium On Usable Privacy and Security (SOUPS) 2011, July 20-22, Pittsburgh, PA, USA. https://dl.acm.org/doi/abs/10.1145/2078827.2078838

Pennycook, G, and Rand, D.G. (2021) "The Psychology of Fake News", [online], *Trends in Cognitive Sciences*, Vol. 25, No. 5. pp. 388-402. https://www.sciencedirect.com/science/article/pii/S1364661321000516

Saberi, A., Vahidi, M. and Bidgoli, M.B. (2007) "Learn To Detect Phishing Scams Using Learning and Ensemble Methods", [online], 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops https://ieeexplore.ieee.org/abstract/document/4427596

Simándi, Sz. (2019) *Közösségi tanulás felnőttkorban Tanulókörök az élethosszig tartó tanulás folyamatában*. Budapest, Akadémiai Kiadó.

Simándi, Sz. – Oszlánczi, T. (2018): "Időskori tanulás - közösségi művelődés". In: Fodorné, T. K. (ed): *Social and economic benefits of university lifelong learning: research, development and innovation*. Debrecen, Hungary: MELLearN Felsőoktatási Hálózat az életen át tartó tanulásért Egyesület pp. 380-384.

Szűts, Z. (2020) "A digitális pedagógia jelenségei és megnyilvánulási formái", [online], *Új Pedagógiai Szemle,* Vol. 70 No.5-6. pp. 14-36. http://upszonline.hu/index.php?article=700506007

Tengely, A. (2001) "Tisza István 1913. januári politikai párbajai" In: Mátyás, B. (ed) *Grastyán Endre Szakkollégium tanulmánykötetet 2*. Pécs, Pécsi Tudományegyetem Grastyán Endre Szakkollégium, pp. 239-276.

Wash, R. (2020) "How Experts Detect Phishing Scam Emails", [online], *Proceedings of the ACM on Human-Computer Interaction*, Vol. 4, No. CSCW, pp. 1-28. https://doi.org/10.1145/3415231