Volume 2, No. 1

2023

ACIG

APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

ISSN 2956-3119
E-ISSN: 2956-4395

Volume 2, No. 1

2023

ACIG

APPLIED
CYBERSECURITY
& INTERNET
GOVERNANCE

NASK

www.acigjournal.com

Printing
ISSN: 2956-3119 | E-ISSN: 2956-4395

Information email: contact@acigjournal.pl

Table of contents

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Letter from the Editor-in-Chief

Corresponding author:
Aleksandra Gasztold, NASK National Research Institute, ul. Kolska 12, 01–045 Warsaw, Poland; E-MAIL: aleksandra.gasztold@acigjournal.pl

ORCID
https://orcid.org/0000-0002-9114-1604

Applied Cybersecurity & Internet Governance
2023;2 (1)

Dear Esteemed Readers,

It is with great pleasure that I introduce the 2023 volume of Applied Cybersecurity & Internet Governance (ACIG). As Editor-in-Chief, I am excited to present to you an array of insightful articles that delve into various facets of cybersecurity and its intersection with governance, technology, and society. At NASK-National Research Institute, the publisher of ACIG, humans are at the heart of technology. As a research institute dedicated to enhancing information and communication networks in Poland, we prioritize research, development, and education to empower users and safeguard the digital landscape, particularly focusing on the advancement of cybersecurity measures and the understanding of cyber threats.

In this issue, we bring together a collection of original research articles that offer diverse perspectives and analyses, catering to readers from academia, the IT sector, policy decision-makers, and beyond. Each article encapsulates cutting-edge research, practical insights, and thought-provoking discussions that contribute significantly to the discourse surrounding cybersecurity and modern online challenges.

Fourteen articles featured in this volume cover a wide range of topics, including supply chain risks in autonomous weapon systems, hybrid warfare, cognitive warfare, cyber resilience at the state level, and the protection of critical infrastructures, among others.

The issue opens with the article “Structured Field Coding and its Applications to National Risk and Cybersecurity Assessments” by William H. Dutton, Ruth Shillair, Louise Axon and Carolin Weisser, which explores the utilization of structured field coding in enhancing national cybersecurity assessments, facilitating cross-national comparative analyses. Moreover, “Predictive Modelling of a Honeypot System Based on a Markov Decision Process and a Partially Observable Markov Decision Process” by Lidong Wang,

Reed Mosher, Patti Duett and Terril Falls, presents innovative approaches to predictive modeling in honeypot systems, crucial for proactive cybersecurity measures. The next paper, titled "Artificial Immune Systems in Local and Network Cybersecurity: An Overview of Intrusion Detection Strategies" by Patryk Widuliński, provides an overview of artificial immune systems in intrusion detection systems, offering insights into recent advancements and future research directions. Complementing this section, the interview held by Rubén Arcos, "Shielding the Spanish Cyberspace: An Interview with Spain's National Cryptologic Centre (CCN)," aims to present the perspective of security institutions involved in monitoring cyberspace for threats.

Furthermore, "Examining Supply Chain Risks in Autonomous Weapon Systems and Artificial Intelligence" by Austin Wyatt delves into the risks associated with AI-enabled autonomous systems, focusing on the vulnerabilities within supply chains responsible for producing such military technologies. In the context of growing threats in cyberspace, Marco Marsili's article, "Guerre à la Carte: Cyber, Information, Cognitive Warfare and the Metaverse," explores the concept of hybrid warfare, particularly within the context of cyber, information, and cognitive hostilities, shedding light on the implications of these phenomena in the modern world. Thereafter, Guillermo Lopez-Rodriguez, Irais Moreno-Lopez and José-Carlos Hernández-Gutiérrez compare cyber attacks on energy infrastructures carried out by Russia and Iran, analyzing their strategies and political implications in the manuscript titled "Cyberwarfare against Critical Infrastructures: Russia and Iran in the Gray Zone." Staying in the area of hybrid warfare in cyberspace, the paper entitled "The Russia-Ukraine Conflict from 2014 to 2023 and the Significance of a Strategic Victory in Cyberspace" by Dominika Dziwisz and Błażej Sajduk examines Russian engagement in cyberspace during the conflict with Ukraine, challenging Western perspectives and discussing Russian cyber warfare strategies.

Moreover, "Tell Me Where You Live and I Will Tell Your P@Ssword: Understanding the Macrosocial Variables Influencing Password's Strength" by Andreanne Bergeron investigates the influence of macrosocial factors on password strength, aiming to offer insights into global cybersecurity policies and configurations. An approach to societal aspects was presented in "Trust Framework on Exploitation of Humans as the Weakest Link in Cybersecurity" by Daudi Morice. This analysis develops a trust framework focusing on the exploitation of human psychology in cyberattacks, highlighting the importance of understanding and mitigating human vulnerabilities in cybersecurity. However, a comprehensive conceptualization of state-level cyber

resilience, offering insights into the capacities required for states to effectively respond to cyber threats is examined by Geoffrey Hubbard in "State-level Cyber Resilience: A Conceptual Framework."

The challenges of the changing digital technology landscape exemplified by efforts in the European Union are presented in two articles. The first, titled "Protection of the EU's Critical Infrastructures: Results and Challenges" by Robert Mikac, analyzes EU legislative acts aimed at improving the resilience of critical infrastructures, focusing on potential weaknesses and suggesting solutions. The next, "Regulating Deep Fakes in the Artificial Intelligence Act" by Mateusz Łabuz, discusses the challenges and limitations in mitigating the negative consequences of deep fake technology. The issue closes with "Creating a Repeatable Nontechnical Skills Curriculum for the University of Southern Maine (USM) Cybersecurity Ambassador Program (CAP)" by Lori L. Sussman and Zachary Leavitt. It presents a case study on the development of a nontechnical skills curriculum for a cybersecurity internship program, aiming to bridge the gap between academia and industry demands in cybersecurity education.

We extend our sincere gratitude to the authors, reviewers, and International Editorial Board whose dedication and expertise have made this volume possible. We hope that the articles presented herein stimulate meaningful discussions and inspire further research in cybersecurity and internet governance.

Thank you for your continued support and readership.

Warm regards,

Aleksandra Gasztold
Editor-in-Chief
Applied Cybersecurity & Internet Governance

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Structured Field Coding and its Applications to National Risk and Cybersecurity Assessments

William H. Dutton | Oxford Martin School, Oxford University, UK, ORCID: 0000-0002-0141-6804
Ruth Shillair | Department of Media & Information Studies, Michigan State University, USA, ORCID: 0000-0003-0341-9096
Louise Axon | Department of Computer Science, Oxford University, UK, ORCID: 0000-0001-5979-7630
Carolin Weisser | Harris Global Cyber Security Capacity Centre, Oxford University, UK

Abstract

Data on cybersecurity capacity building efforts is critical to improving cybersecurity at national levels. Policy should be informed not only by measures that allow internal assessment of strengths and weaknesses that enable cross-national comparisons. The International Telecommunications Union (ITU) and its Global Cybersecurity Index (GCI) has used a standardized survey that has been adapted and used in multiple national assessments by the Global Cyber Security Capacity Centre. This adaptation includes an addition of open field coding assessments that rely heavily on trained experts and interactions with national focus groups. These assessments are checked using multiple coders to increase reliability and reduce bias. This process of 'structured field coding' (SFC) is an approach to collecting and coding observations based on multiple methods, quantitative as well as qualitative. This approach differs from open field coding in providing a set structure for coding observations from the field based on established frameworks for assessment. The SFC process is explained along with a discussion of the origin and the advantages and limitations of this methodological

Received: 12.06.2023

Accepted: 18.10.2023

Published: 27.10.2023

Cite this article as:
W.H. Dutton, R. Shillair, L. Axon, C.W. Harris "Structured Field Coding and its Applications to National Risk and Cybersecurity Assessments," ACIG, vol. 2, no. 1, 2023, DOI: 10.60097/ACIG/162857

Corresponding author:
Ruth Shillair, Department of Media & Information Studies, Michigan State University, USA, ORCID: 0000-0003-0341-9096; E-MAIL: shillai7@msu.edu

approach. It can be used in a variety of studies but is presented here as a means to integrate data for cross-national comparative analyses. Its application to improving the reliability and validity of data collection across a region, such as the EU, would help stakeholders evaluate where they should invest resources to improve their cybersecurity capacity.

Keywords
cybersecurity capacity building; structured field coding; cybersecurity analysis; multi-methods security research

1. Introduction

Efforts to build a nation's capacity to withstand cyberattacks and other risks to cybersecurity contribute to a nation's security and economic vitality. Almost all our modern systems: communication, economic transactions, record keeping, and critical infrastructure are controlled through computerized systems. These systems increase efficiency and speed, yet simultaneously they can be an Achilles heel, as each system introduces additional surfaces that are vulnerable to attack. Thus, it is important to build an ecosystem that is robust and resilient to cyberattacks. This process of cybersecurity capacity building is systematic, touching many societal dimensions [1]. The investment in cybersecurity capacity building pays back in the function of critical infrastructure and economic vitality [1, 2] as well as new legal and policy frameworks. While efforts to proactively address security problems seem intuitively valuable, they are new, meaning there is relatively little research on whether they achieve their intended objectives. This paper takes a cross-national comparative approach to determine whether there is empirical support for investing in capacity-building. It reflects field research from 73 nations as well as comparative data analysis. These efforts recognize that improved cybersecurity capacity is a multi-dimensional effort, it includes not just technological improvements, but also improvements in education, awareness, and training [3].

The first step in capacity building efforts and cybersecurity policy initiatives is often establishing basic "norms" as to what are best practices across these dimensions [4]. Both regional [5] and international efforts have worked to develop norms based on input from interdisciplinary teams. One of the larger efforts is the International Telecommunications Union's (ITU) Global Cybersecurity Index (GCI) [6]. The GCI uses questionnaires and expert advice to rate countries on legal, technical, organizational strategies and plans, levels of

international cooperation, and other capacity building measures to create an overall ranking. A similar National Cyber Security Index (NCSI) has been developed by the non-profit consultancy organization, the e-Governance Academy Foundation, based in Estonia as a joint initiative of the Government of Estonia, the Open Society Institute, and the United Nations Development Program [7]. The NCSI is one project in its cybersecurity program of activities, which includes consulting projects for various nations along with its development of an index for a growing number of countries [7]. The index "measures the preparedness of countries to prevent cyber threats and manage cyber incidents" to compose a database that is publicly available [8]. The evidence is either provided by the nation's government officials, an organization or individual, or by the NCSI team through desk research on legal acts, official documents, and official websites. These sources are used by an expert group at NCSI to make summary assessments on multiple aspects for each nation [8]. In such ways, the selection of index items is a dynamic process that is founded on a solid understanding of published research, comparison to similar studies, and adaptation to emerging trends.

Once index items and norms of assessment are established the next challenge is the design of a methodological process. The research process should assess the performance of nations over time and allow comparison with other nations. Evidence is often based on indicators drawing from a multitude of sources, ranging from different institutions and departments as well as different methods, such as in-depth interviews, questionnaires and surveys, and the aggregation and interpretation of data collected for other purposes, such as national census records. Not only is the evidence collected from multiple sources, but also the outcomes of the assessments are critical to multiple stakeholders, each of whom have a strategic interest in how different sectors or nations are rated. There are also the challenges of assuring norms assessment is applicable to countries at different stages of economic and technical development, particularly as those with less experience and centrality of Internet use may be more vulnerable to cyberattacks [9, 10]. Furthermore, there are concerns when developed nations are assisting developing nations in cybersecurity development, as these efforts might be construed as supporting the geopolitical interests of developed nations [11], which could be branded as digital neo-colonialism rather than a win-win strategy. Thus, there is a need to develop tools that can be readily adopted by and scaled to small or large countries, validated and employed by the adopting nation, and yet reliable and standardized sufficiently to be comparable across nations for accurate and actionable insights.

Given limited resources, and many sources and types of data within each nation, the question is: How national assessments be done efficiently and in a reliable and valid manner that can be replicated and compared over time and with other nations? Structured field coding is an answer to this question. It is an approach which increases objectivity in cybersecurity capacity building assessments. Given its quantitative basis, it can also be linked with related data, such as from external risk assessments.

Every approach to the measurement of cybersecurity capacity building efforts of a nation or region has strengths and weaknesses. Structured field coding is offered as a means for addressing some of the key problems with developing reliable and valid indicators that are comparable across nations and over time. This paper explains, illustrates, and critically examines the concept of Structured Field Coding (SFC) and discusses limitations on its use.

1.1. Approaches to National Case Studies and Comparative Research

One of the first steps in measuring cybersecurity capacity building it to establish the basic parameters that are needed for capacity building. In the area of assessing the maturity of cybersecurity capacity building. There have been many approaches for developing comparable assessments. Many widely used scales have been developed across multiple societal dimensions that impact cybersecurity capacity, either directly or indirectly. These often include such aspects as technical norms, educational programs, legal protocols, and policy mandates. The ITU's Global Cybersecurity Index, is one example, along with a NIST cyber security framework [12], but these are only two of many other approaches that have been developed that overlap in their methodology and empirical indicators.

This paper focuses on one approach where structured field coding has been developed. Oxford University's Global Cyber Security Capacity Centre (GCSCC) has developed, in collaboration with over two hundred international experts, a Cybersecurity Capacity Maturity Model for Nations (CMM) and an approach to assessing nations that has been deployed in over 80 nations to date.

The CMM reviews cybersecurity capacity across five dimensions: (1) Cybersecurity Policy and Strategy; (2) Cyber Culture and Society; (3) Cybersecurity Education, Training, and Skills; (4) Legal and Regulatory Frameworks; and (5) Standards, Organizations, and Technologies. Each dimension consists of a range of factors that describe what

it means to possess cybersecurity capacity in that dimension, and aspects of each factor that enhance maturity. A set of indicators for each aspect of those factors is used to gauge cybersecurity maturity along a five-stage spectrum, ranging from (1) start-up; (2) formative; (3) established; (4) strategic; to (5) dynamic [13].

During the initial years of deploying the CMM, data-gathering involved in-country stakeholder consultation (typically 2–3 research staff visit over the course of three days), complemented remotely through desk research. The in-country consultations, which relied mainly on modified focus groups, was to yield evidence for assessing capacity building for each nation in ways that can be used both to recommend capacity-building initiatives for nations but also for making comparisons across nations. But as in the case of national comparisons, the regional assessments would ideally be comparable across multiple units to estimate capacity levels across the nation.

Investment, and policy decisions are inevitably made based on national assessments, whether these are limited to mere hearsay or anchored in systematically empirical and accountable evidence (Box 1). Thus, a rigorous and accurate assessment is advantageous for data-based decisions. Nations cannot avoid being challenged for their policy decisions, but the more reliable and valid the evidence is judged to be, the easier it is to demonstrate the foundations for prioritizing areas for investment.

VALIDITY: concerns the degree that an indicator is measuring what it is intended to measure. Are you measuring what you think you are measuring? Have there been multiple tests and expert input to support validity? Does cybersecurity capacity maturity indicate the resilience and status of a nation in responding to breaches and other attacks on cybersecurity?

RELIABILITY: refers to the degree that an indicator can replicate an underlying trait accurately or consistently. Will an approach to capturing a national level of cybersecurity maturity be capable of yielding the same results if replicated, such as by a different team of researchers?

Box 1. Reliability and Validity

1.1.1. Aggregate Data Collection

One approach that is common in relatively well-defined areas is to combine new or existing aggregate national indicators to assess performance, such as in the areas of economic development or freedom of the press. Cross-national comparative research is often based on field research or the use of available aggregate data that might have been collected for other purposes, but which can be used to extract empirical indicators of national similarities and differences.

For example, Freedom House rates countries or territories on 10 indicators of political rights, such as free and fair elections, and 15 indicators of civil liberties, including the rule of law, that are each ranked from 0–4, where 0 represents the lowest level of freedom and 4 the highest. In 2020, Norway and Finland were ranked 1 and 2 on their respective scores on the press freedom index [14]. Other aggregate data approaches have been used to develop indicators of governance [15], and cyber power [16], for example. Many of these kinds of studies or rankings are done by a single organization following developments and activities across multiple nations across the world. Aggregate data can be drawn from research conducted by other organizations for other purposes. The benefit of aggregate data is that it is often collected by well-funded and highly respected organizations and can be used to empirically test concepts that otherwise couldn't be tested at a large scale [2].

The use of the same data for multiple studies is almost demanded by the time and cost of developing national indicators. The challenge with aggregate data is that the data items collected limit the research questions that can be addressed. Thus, aggregate analysis is limited to the relationships between verified data in the sets under consideration, requiring some collection of original data.

1.1.2. Field Research and Data Collection

Original collection of data for national assessments allows for collection of more than just standard measures, it allows for customization and contextualization of the data collection tools to fit the specific country and phenomenon being studied. As in many domains, understanding the unique needs data sources within a country often takes time and expertise. For example, most national comparative research in the study of government and politics has been conducted by individuals who lived and/or worked for a sustained period in a particular nation other than their own. They had become expert participant-observers of activities in the nation that is their object of

study and most often develop their findings. A classic demonstration of this method is Alexis de Tocqueville's examination of democracy in America in the early 1800s (N = 1), or a comparative case study (N = 2 or more) [17]. A more recent example is Gabriel Almond and Sidney Verba's study of political attitudes in the United States, Great Britain, Germany, Italy, and Mexico, which drew on qualitative observations as well as comparative survey research data collection [18]. The World Internet Project (WIP)[1] collects data from a growing number of countries, focused on issues of the digital divide.

1 — https://www.worldinternetproject.com/

Given the intense commitment of time, labor, and expertise, it is rare for field research about cybersecurity to be conducted in many nations. This work has often been limited by the ability of individuals or a small team to be directly involved in observing, interviewing, and comparing nations. Even though this method can produce high quality reviews with many actionable insights, they can be prohibitively expensive and require commitment from many stakeholders.

Additionally, the need for strategies that might enable larger and more distributed teams to gather comparable data in the field, new methods are being developed that include the use of interviews and participant-observation. These have the potential to give deep insights to help guide policy changes that are needed in this dynamic domain. However, to develop these methods, that would allow both standardization for comparative value and flexibility to adjust to changing threats; it is important to first examine what has already been used in other leading cybersecurity capacity assessments.

1.1.3. National Risk Assessments and the UK's National Cyber Risk Assessment (NCRA)

National risk assessments are focused on identifying major risks facing a nation, developing estimates of how likely it is that the nation will experience each risk, and estimating the severity of the risk. Given the uncountable number of known and unknown risks that nations might face, even in a constrained area, such as cyber, these assessments rely heavily on the judgements of subject matter experts in a wide range of sectors within each nation. The Organization for Economic Co-operation and development (OECD) and other major international organizations view risk assessment as critical to managing these risks and achieving national economic and social goals [19].

The government of the United Kingdom has developed a Cyber assessment Framework from the National Cyber Security Centre

that has four objectives: managing security risk, protecting against cyberattack, detecting cyber security events, and minimizing the impact of cyber security incidents [20]. Each of these items are broken down into measurable items that focus on IT systems and network solidarity [20].

A cyber risk assessment captures the judgmental ratings or forecasts of relevant experts on the priority, likelihood, and level of cyber risks across many sectors and critical infrastructures of specific nations (Box 2). While the questions can be structured in similar ways cross-nationally, the answers are sufficiently unique to each nation that comparison is difficult. They are developed to help nations better anticipate and manage potential risks to their nation and not designed with an aim of cross-national comparison. As with cybersecurity capacity assessments, it is difficult to study such assessments across multiple nations in comparable ways.

National transportation and telecommunication systems, including the information and communication technologies (ICTs) that support them, are two huge critical infrastructure sectors. Depending on the level of analysis, the number of CIs in focus vary from four to nearly twenty viewed as necessary to the functioning of the nation. These are the systems, networks and assets that are essential to the functioning of a society [21].

Box 2. Critical Infrastructure Sectors.

Discussions between those involved in research on cybersecurity maturity and risk assessments identified ways to improve approaches to each area of research, also explored the potential efficiencies and synergies of integrating these two heretofore separate activities. The idea of an integrated cybersecurity maturity and risk assessment (Cyber-MRA) is attractive, creating one unified assessment to give insights and guide policy decisions. However, given the different organizations, traditions, and methods tied to each, their integration is challenging. Assessments that rely on qualitative analysis, focus groups, and interviews often produce deep insights, but these are not always quantifiable and comparable across sessions of data gathering. On the other hand, field surveys with strict "fill in the bubble" type approaches yield problematic data even though they allow quantitative analysis and comparison across data sets. A potentially strategic innovation for improving on and integrating these two approaches involves the use of what we have called "structured field

coding". The next section describes this approach and then moves to a discussion of how it can enhance and potentially integrate both maturity and risk assessments.

1.2. Structured Field Coding (SFC)

Structured field coding (SFC) can potentially advance the study of national cybersecurity maturity and national risk assessments as well as provide a means for better integrating multiple methods and indicators involved in both qualitative and quantitative approaches. The following is the history of the method, its growth and how it has been utilized in national cybersecurity assessments to provide rich insights and actionable items for building cybersecurity capacity.

1.2.1. The Origins of SFC

SFC was invented to solve a set of problems that arose in the study of the early use of computing in forty US cities in the late 1970s [22]. The study, entitled An Evaluation of Urban Information Systems (URBIS), was one of the first systematic studies of the use and implications of computer systems in American local governments. It was funded by the US National Science Foundation and based on what was then the Public Policy Research Organization (PPRO) at the University of California, Irvine. The principal investigator was Professor Kenneth Kraemer, who led the team of co-principal investigators, including James Danziger, William Dutton, Rob Kling, and Alex Mood. The study began with a survey of all 403 US cities with populations over 50,000. In 1975, circa the time of this study, this was about the size at which a city might have had one or more computers and associated applications, although nearly a quarter (23% or 93) of the cities over 50,000 inhabitants had such a negligible level of computing that they were dropped from the study.

The URBIS team devised a means for stratifying all US cities on key policy variables, which involved using standard demographic data to estimate scores on indicators of computer use for those cities for which data was not available. The team then randomly sampled cities such as to maximize variation on the major variables of interest to the study, such as the centralization or decentralization of computer facilities. This approach was called the "future cities research design" [23].

This analysis led to the selection of a stratified random sample of 40 US cities that were then studied more intensively. The research

formed the basis of numerous publications across all the investigators including two major academic books [22, 24], which together provide considerable evidence of the academic merit and acceptance of the study's methodology.

1.2.2. The Invention of SFC

It was in the pilot stages of this in-depth comparative study of the 40 sampled cities that methodological challenges began to arise. First, there were multiple teams going into the field, most composed of two researchers spending about two weeks in each city conducting interviews, visiting departments, and observing work around several foci of the study, such as in the use of computer-based data in policy analysis, detective investigative support, and other applications representative of different "information processing tasks". In pretesting our approach through a small set of "mini-cases", it was apparent that different investigators tended to confirm their preconceived notions, demonstrating the issue of researcher bias. This is a long known issue, even de Tocqueville is said to have had preconceived views on America, as one French critic said: "He had thought it all out before he learned anything about it [America]" [25]. This led to developing ways to make the research more objective versus relying too heavily on any one person's preconception, or other subjective or judgmental rating, and also to ensure multiple points of view and greater accountability – creating an ability to double-check the conclusions of those who did the field work.

Reducing researcher bias was accomplished using SFC. Essentially, each of the two or more researchers in the field would answer the same questions the entire team had considered critical to the study. They could use in-depth interviews, observations in the field, desk research, and informal discussions with staff and politicians, for example, to arrive at their answers. The team would then compare and contrast their answers and resolve differences of opinion across the members and explain in notes why the city was coded in the way it was finally determined. This method reduced the impact of bias from any particular researcher, while providing a means for integrating multiple observations into a single code. It also made every code more credible since it was the product of multiple observations and explanations for why a city was coded as it was, provided by those individuals who were observers in the field. Thus, it was a structured way to establish "inter-coder reliability" across data that was collected through multiple methods.

2. Methods: SFC in action

Structured field coding (SFC) refers to the development of predetermined questions and potential responses that are answered by the researchers while they are in the field. This is similar to a survey, but different in that the researcher completes the items based on evidence from the field while the data is being collected. Responses can be refined but are initially coded while fresh from interviews and observations in the field context. Using desk research, discussion group transcripts, and interviews with those informed about particular topics, the researchers in the field aim to be in a position to answer each question. The answers to each question are then used to operationally define each indicator. Supporting data is then available to others in the research team to test items for reliability and objectivity.

2.1. The URBIS Example

In the URBIS study, for example, a key question concerned whether the use of computing in cities would shift influence or power to one or another kind of actor [23]. One item asked about the use of data in the city: "In general, has the use or design of data banks [a term of the 1970s], their analysis or the distribution of findings tended to shift relatively more influence away from or to any of the following: [Manager or chief administrative officer; Mayor and staff; Council and staff; Departments; and Data bank custodians], with the following response categories: "Given less influence to; No discernable shift; Given more influence to" [23, p. 200]. Researchers were asked to provide notes to defend their response. A related question was: "Are various computer-based reports and special analyses generated from operational data used in responding to individual or citizen group requests or complaints?". Coding was provided for the manager, mayor, and council, with each coded separately with the following response categories: Cases cited that it could have been but was not used; Believe it is not used; Undecided, mixed; Believe it is used, Cases cited that it is used [23, p. 201]. This exact question was addressed in an interview one researcher had with the city manager of a large city with a city manager form of government. The manager said such analyses were never used, but the researcher went with him to the council meeting that followed, when the manager's office presented the results of systematic modeling where the data in fact were being used. In such ways, multiple sources are reconciled to arrive at a researcher's coding from observations and data collected in the field based on a pre-designed set of codes.

Another innovation tied to the use of SFC in URBIS was the identification of a set of information processing tasks (IPTs) that not only

indicated the levels of data use, but also flagged areas that could be studied in more depth. In the late – 1970s, the application of computing in governments ranged across at least six IPTs (Box 3). For example, by looking at a specific task, such as record searching in supporting detective investigations, or record-keeping in traffic ticket processing, it was possible to have a more concrete empirical basis for assessing the impact of computing. Looking across these IPTs within one city provided an overall picture that was more grounded in the government's use of computing [22]. It is possible that risk assessments might identify a set of comparable CIs that could be a focus for more concrete and detailed analysis in an analogous way.

1. Record-keeping, such as in traffic ticket processing
2. Calculating/printing, such as in budget control
3. Record-searching, such as in detective investigative support
4. Record restructuring, such as in policy analysis
5. Sophisticated analytics, such as in police patrol manpower allocation
6. Process control, such as in budget monitoring and control

Box 3. Information Processing Tasks Defined by URBIS.

2.2. The CMM Example

In the context of the CMM assessment, for example, there is a question with responses for whether a nation has a cybersecurity strategy (Box 3). Focus groups often have responses much like the URBIS experience of several decades ago. Many are unaware of efforts going on in other departments. Additionally, even experts seeking to find evidence of operationalizations in action might not find all the same evidence. If two or more researchers go to a country, they would code responses to this question, and together, they would reconcile any differences in order that they agree on the best code for the country, based on what they have learned. They assign a quantitative number to their qualitative judgements. In instances when they have fundamental differences, which cannot be easily resolved, they would go back into the field via an email, conference call, or use desk research to resolve the differences. For example, they might have been told a strategy was in development, but discover a published report online, and switch their coding to "4" for "yes, and it has been published".

The advantage of the SFC method is its ability to capture data that is not obvious, as well as items across domains, which might be missed

from relying solely on a survey of participants. Cybersecurity capacity building is a multi-faceted, multi-disciplinary effort that requires advanced application of expert knowledge. Measuring progress to evaluate efficacies is a complex, yet critical effort. Just as the example of the URBIS application, the SFC allows stakeholders to see where they are not utilizing resources available to them to improve policies and processes.

Q1. Does the country have a national cybersecurity strategy? (Circle Response)

5... Yes, the country's strategy has been cited as "world leading"
4... Yes, and it has been published
3... Yes, but not published
2... No, but it is in development with a draft or outline
1... No, but the processes for strategy development have been initiated
0... No, strategy does not exist

Evidence, Examples:

Box 4. A Question from the CMM

2.3. A Critical Perspective on the Approach

SFC is a relatively simple idea that provides a flexible approach that increases research validity and reliability when dealing with a complex array of methodological approaches. This is especially important when researching cybersecurity capacity building efforts as it touches on technological, educational, legal, communications, and societal domains.

2.3.1. Strengths of SFC

1. *Reliability of Multiple Observations and Codes:* SFC embeds the use of multiple observers, and coders for each question. This is designed to enhance the reliability of the code agreed across researchers. As differences that emerge across coders will lead to notes and explanations of how the code was resolved, the notes also enhance the reliability attributed to the resulting data.

2. *Areas of Uncertainty or Lack of Awareness.* Capturing multiple (independent) observations at the indicator level also enables

the researchers to identify the specific indicators on which there was uncertainty (on which two or more observers significantly disagreed, for example), and which therefore need continued attention, such as in follow-up calls to the field work, in order to resolve. In many online or remote collection of data these discrepancies would not be known. Discrepancies across coders might also be interesting evidence of conflicting viewpoints in the country or differences in the knowledge base of the participants interviewed. In the example of the CMM, this method helps paint an even richer picture of cybersecurity maturity in the country and identify problematic or uncertain areas of capacity building.

3. *Capturing Detail at the Indicator Level.* For the CMM, as one example, each question represents an indicator of one or more aspects of maturity. Having an agreed coding for each indicator provides more variation and evidence at the indicator level for each nation. Individual indicators can then be used alone or in some descriptive and comparative analyses, as well in calculating maturity levels of their respective aspects.

4. *Evolvability of Operational Definitions.* Coding at the indicator level makes it possible to refine and revise any operational definition of any variable including each respective question. In the case of the CMM, the model can be evolved simply by revising the operational definition of aspects, such as moving an indicator to be grouped or combined with a different aspect. In addition, the team can operationally define why a country is given a particular maturity code on any given aspect. Since each indicator related to a maturity code for any given aspect is recorded, any change in the definition of an aspect can be accommodated by changing the operational definition – how different indicators are combined. In defining an aspect in a new way, using the existing indicators, the new aspect can be quickly recalculated to obtain a new maturity score defined by the model. Researchers can draw from the existing indicators. That is, the model can evolve, and the existing indicators can be used to recalibrate maturity levels.

5. *Precision of Comparisons Over Time and Cross-Nationally.* The operational definition of indicators and maturity or risk levels will enable more reliable and operationally defined variance across countries, and more reliable and valid measurements of maturity or risk over time and cross-nationally. For instance, by relying primarily or only on modified focus groups, the researchers make judgmental ratings of the maturity levels of each aspect given the

observation of particular „indications" of their level. But this does not capture variation on the actual indicator – only that it might have been observed. More precise, operational indicators would enable less dependence on judgmental ratings and better able to capture minor differences cross-nationally and longitudinally. For instance, in the example question above, you might find that a country had published its strategic plan, another which has not. So small variations would be more visible and subject to analysis.

6. *Transparency and Accountability.* SFC leads to national ratings that are more transparent and accountable as anyone could see and question the operational definitions of ratings, and the indicators used. So SFC would enhance transparency of the data and analyses based on cross – national or over-time comparisons.

7. *Integration of Data from Multiple Sources.* One of the most valuable advantages is that by enabling the use of multiple data sources in coding, it is possible to draw not only from multiple data on the same nation but also multiple studies, if conducted concurrently by the same or even different research teams, if they used the same SFC. A later section of this paper will illustrate its potential for integrating the study of national cybersecurity capacity building with the study of national cyber risks.

8. *Integration across Sectors or Infrastructures.* Finally, many studies of governments or nations cannot study all activities, sectors, or infrastructures. A pragmatic but also a valid approach is to identify a sample of individuals, departments, sectors, or infrastructures to study in more depth, but in ways that can be compared and/or aggregated to a higher level of analysis. SFC could be developed to ask similar questions about different objects of analysis in ways that the answers are more comparable and less problematic to aggregate.

2.3.2. Limitations of SFC

There are weaknesses or limitations of SFC – it is not a silver bullet for resolving major challenges in national and cross-national comparative research. These include:

1. *Limits on Independence of Coders.* Each indicator defined by a maturity model or risk assessment should be coded by two or more independent observers. With at least two individual researchers going into the field there is nevertheless the likelihood of some bias of individuals to confirm their

preconceptions – a confirmatory bias – but also the potential for interpersonal influence to undermine the independence of the coding. Of course, at the end of the day, researchers need to compare codes and resolve differences of opinion, which demands some role of interpersonal influence and compromise. However, these sessions can be conducted in light of concerns over avoiding any confirmatory and group think biases, and most other research approaches face the same threats, such as how different individuals code group discussions

The potential for any lack of independence is addressed in several ways. First, desk research is likely to involve researchers beyond the field team. Secondly, the explanations of codes could indicate a lack of independence, which would be apparent to those beyond the field team drafting the report. Finally, the codes and the report based on them will be reviewed by experts outside the field team, including experts within the respective nations. Judgmental ratings and SFC will be sufficiently transparent that they will be subject to several stages of accountability.

2. *Time Demands of Coding.* In some respects, the use of two independent coders might be viewed as doubling the workload on the research team, but this is a compromise that will lead to more reliable and valid indicators (Box 1). As discussed in the section on the origins of SFC, the risks of a single coder appear greater than threats that two coders, but one more coder will not eliminate such a risk. Nevertheless, clarifications of codes by two coders will add more texture to the meaning of the code and the evidence behind it.

3. *Pressure to Reduce the Number of Indicators.* Time demands do exert pressure on the study team to minimize the number of questions or indicators included in the study. It is a natural expectation that subject matter experts in cybersecurity or cyber risk will want to be as comprehensive as possible and include every conceivably important question. However, there reaches a point when the time required in the field surpasses that allotted, which threatens the care and precision of the coding process. The research team needs to include enough indicators to get a reliable estimate of aspects related to those indicators but avoid temptations to be comprehensive. It is easy to write questions but difficult to answer and code them. This creates an inherent problem with the team creating too many questions in ways that inadvertently reduce the quality of the research. A survey has limits imposed by the time that respondents are willing to spend answering questions. This

places severe constraints on the number of questions asked in surveys, for example. Likewise, a specified time in the field places similar constraints on the number of interviews, discussion groups, and participant – observation that is possible in a single nation. The research team must therefore exercise considerable discipline in reducing redundancy, tangential questions, and exceedingly complex coding issues to ensure that the field research is completed. Just as a set of survey questions does not need to be comprehensive to provide an indication of a behavioural or attitudinal propensity, neither do the indicators included for SFC need to be comprehensive. What indicators are necessary to make a judgement on the relative maturity of a nation in a particular area of cybersecurity?

4. *Risk of Failing to Gain Multiple Codes.* It is possible a researcher might fail to get evidence about all indicators, so two observers will enhance the likelihood of at least one researcher collecting evidence from interviews or observations that can be used for coding the indicator. This is a pragmatic reality of field research. It is not ideal, but the effort would strive to obtain evidence from each researcher on each question, recognizing that this will not always be possible in the time allotted to field research, and the strategies for gathering data in the field, such as in dividing in-depth expert interviews up between the two or more researchers. Desk research and post-field research interviews, such a via video conferencing, can be used to address any doubts raised by the lack of double coding.

5. *Limits to the Detail and Precision of Rapid Field Research.* Surveys are blunt instruments, seldom capable of capturing the precise level of detail many journalists, public officials, and other subject matter experts expect from them. Likewise, any research based on multi-methods conducted over a very short period of time – such as a few days – cannot be expected to be as precise as one would wish. For example, any data collected today, might be different tomorrow. Any evidence uncovered by two researchers over 2–3 days might miss additional evidence that could have affected the coding of indicators. Any period of time chosen for the research might be influenced by events in the national context or even the personal situation of the person interviewed that would bias the observations, such as a change in administration. These limitations need to be recognized and efforts should be made to identify any problematic data, but the team also needs to realize that they cannot be totally overcome.

6. *Unknown Knowns.* By having a pre-defined "structure", or sets of questions and codes, the study could fail to capture information that is pertinent to cybersecurity maturity assessment that was not already identified and added to the SFC's codes. In the case of the CMM study, this is addressed in part by continuing revisions of the underlying model, based on lessons learned from past assessments and new technical and legal approaches. But SFC does lean heavily on the research team having the right pre – conceived notions of what are the best indicators of cybersecurity maturity, and not missing any key developments. However, discussions gained in in-depth interviews and modified-focus groups are recorded and can inform each case but would be less likely to be valuable for comparative study. Employing a more "grounded theory" approach, where the participant-observation and related data are collected in a more open-ended manner [26], would focus more attention on researchers reviewing the interviews, notes and discussions in an iterative manner to identify the codes to be applied to it. That said, this approach is still framed by the less structured ideas that frame the questions and observations of the researchers and tend to develop more unique frameworks for each case study that could be compared cross-nationally but in different, broader, and more thematic ways. Moreover, the use of SFC adds numbers to qualitative data. It does not erase or substitute for qualitative and other quantitative data and observations. What it does do is insist that the researchers cover areas defined by the SFC and in this respect it steers data and observations in ways that might not be incorporated in a more open – ended approach to following the evidence.

2.4. SFC Enhanced by Modified Focus Groups

The use of structured field coding (SFC) could complement and augment the use of modified-focus groups for field research on cybersecurity capacity building, or the use of discussion groups focused on cyber risks. Past CMM reviews relied greatly on what the GCSCC team has called modified – focus groups, which have several limitations that can be reduced using SFC.

The term "modified-focus groups" (MFG) is meant to convey the divergence of this approach from traditional focus group methods per se, a process invented by a famous sociologist, Robert Merton, in the 1950s to study opinion formation, such as study of why people support a policy. Standard approaches to focus groups are generally used to surface a wide range of opinions, through open-ended

questions, such as seeking to understand what people understand by the concept of cybersecurity or cyber risks. They are excellent approaches for understanding how to design a questionnaire, for example, a focus group discussion of how the government thinks about cybersecurity capacity might help us design more structured questions to which individuals could respond.

However, standard focus groups are not designed to reach a consensus on a question or an issue, but to foster a range of opinions. The exact opposite aim is the normal rationale for a MFG. MFGs are designed to elicit a range of opinions that lead to some consensus, such as whether a nation follows a particular practice. Moreover, MFGs bringing together individuals from government, business and industry, civil society, and academia violate some assumptions that underpin the value of collective intelligence. MFGs can bring in rich insights beyond the more objective SFC and even though it is a challenge to combine the data, the process has led to success in the CMM national evaluations based on such criteria as construct validity – judged by empirical relationships with other indicators expected to be associated with the indicator being measured.

One of the challenges of MFG it that they are also very difficult to validly replicate. If a specific field researcher moderating the discussion chose to kick off discussions with their own inspired prompts and questions, based on the specific context and informal discussions, it could skew the entire group. Each focus group could be primed somewhat differently and would therefore possibly react to somewhat different sets of questions and prompts. They can be replicated only in the broadest sense of doing multiple focus groups, with each likely to be composed of different sets of individuals and with the likelihood of being primed by the early statements and questions raised by the participants and moderator. Thus, despite such challenges, MFGs work well enough, as reflected in the face validity and construct validity of measurements, while also playing an important role in awareness raising and networking. In addition, the MFGs provide important indications of the state of knowledge in the country and where knowledge gaps might be, which can be important to influencing practices. It can be important to gain a sense of whether people from a range of sectors are generally aware of various strategies, legislation, activities (e.g., awareness-raising activities, to understand whether they have a good reach) – and not just to know whether these things exist by asking the experts. For example, if awareness-raising activities exist, but very few people know about them, this could explain any lack of success.

Results

Structured field coding is supported by multi-method, multi-sourced data collection, thus seeks to increase validity and reliability in critical research. Figure 1 illustrates how multiple data sources can feed into judgements made on the coding of a nation on any number of criteria. SFC would be used to code each indicator of both studies and in doing so, it would convert data from any source into a comparable national indicator. That said, some data would not need SFC, such as the population of the nation and other demographic indicators that would directly fall into the data set if they are at the national level of analysis. Since major approaches to cybersecurity capacity assessments and cyber risk assessments use many of the same data sources, it is feasible to integrate the conduct of both assessments to create an integrated national data file (IND).

An example of how operationalization measures of indicators would work is given in the following examples. Firstly, an analysis of an indicator of the quality of cybersecurity education in a nation. This might be feasible to gauge through desk research using existing reports, news, and the web and related social media. On the other hand, an indicator of Internet use in each respective nation could be measured through existing surveys [6], or bespoke surveys created by the study team. Rating the indicator of the risk of cyber-attacks, along with their likelihood and severity, could be gained through NCRA surveys, governmental, business and industry reports, and all followed up with expert interviews. The multiple sources, collected through multiple methods pool together and strengthen insights.

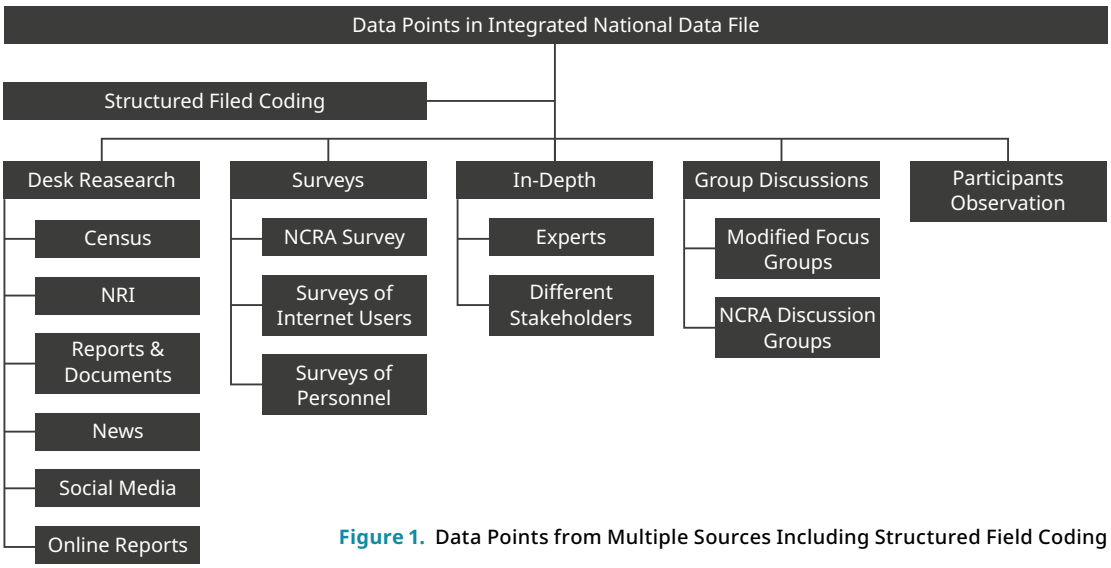

Figure 1. Data Points from Multiple Sources Including Structured Field Coding

3. Discussion

The development of quantitative data from qualitative research is a challenge in the best of circumstances. However, using SFC as part of a holistic data collection process, is a challenging process as it takes funding of trained researchers and a long term commitment to support the required levels of data collection and analysis.

Structured field coding, even though it has been foundational in measuring cybersecurity capacity maturity across many nations through various programs, has to potential to do even more as it is scalable and comparable. A complete assessment relies on multiple methods of data collection. Additionally, this technique could also be used to integrate the data collection process in ways that reduce duplication (each assessment has some common indicators, such as demographics) and create an integrated national data (IND) file that would facilitate analysis of the relationships between aspects of cybersecurity and capacity building. Policy measures that encourage the use of robust measures such as SFC allow nations to measure progress in their capacity building efforts. It is possible to maximize reliable variance across nations in ways that would better support cross-national and longitudinal analysis. These types of analysis are essential to better understand the impacts of less direct measures of capacity building (e.g., legal changes or educational efforts) impact long term outcomes.

4. Conclusions

Structured field coding (SFC) provides a robust technique for reducing redundancy while enhancing the efficiency and effectiveness of cross-national comparative studies. It provides a structured way to enhance inter-coder reliability across data collected through multiple methods. At the same time, does not lose any of the virtues of multiple methods, such as focus groups or in-depth interviews. And it allows the research to amalgamate data in a documented and transparent way across multiple methods to move into a simple structured frame. A promising potential application is the integration of cybersecurity maturity assessments done in conjunction with cyber risk assessments. The resulting integrated national data file would be more powerful than either one data file on its own in supporting a nation's self-assessment and help bring together a wide range of analytical approaches to key questions.

Funding

This work was supported by the Oxford Martin School | Global Cyber Security Capacity Building Center https://gcscc.ox.ac.uk/home-page and their sponsors.

References

[1] S. Creese, W. H. Dutton, P. Esteve-González, R. Shillair, "Cybersecurity capacity-building: cross-national benefits and international divides, *Journal of Cyber Policy*, vol. 6, no. 2, pp. 214–235, 2021, doi: 10.1080/23738871.2021.1979617.

[2] W. H. Dutton, S. Creese, R. Shillair, M. Bada, "Cybersecurity Capacity: Does It Matter?, " *Journal of Information Policy*, vol. 9, pp. 280–306, 2019, doi: 10.5325/jinfopoli.9.2019.0280.

[3] R. Shillair, P. Esteve-González, W. H. Dutton, S. Creese, E. Nagyfejeo, B. von Solms, "Cybersecurity education, awareness raising, and training initiatives: National level evidence-based results, challenges, and promise," *Computers & Security*, vol. 119, p. 102756, 2022, doi: 10.1016/j.cose.2022.102756.

[4] R. Collett, "Understanding cybersecurity capacity building and its relationship to norms and confidence building measures," *Journal of Cyber Policy*, vol. 6, no. 3, pp. 298–317, 2021, doi: 10.1080/23738871.2021.1948582.

[5] M. Górka, "The Cybersecurity Strategy of the Visegrad Group Countries, "*Politics in Central Europe*, vol. 14, no. 2, pp. 75–98, 2018, doi: 10.2478/pce-2018-0010.

[6] ITU, "Global Cybersecurity Index." 2018. [Online]. Available: https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2018-PDF-E.pdf. [Accessed: Aug. 8, 2023].

[7] EGA 20, "The National Cyber Security Index Ranks 150+ Countries' Cyber Security Status." [Online]. Available: https://ega.ee/news/national-cyber-security-index-ranks-150-countries/ [Accessed: Nov. 11, 2023].

[8] NCSI, "National Cyber Security Index Methodology 3.0," 2023. [Online]. Available: https://ega.ee/wp-content/uploads/2023/08/NCSI-3.0_Methodology.pdf. [Accessed: Aug. 9, 2023].

[9] L. Muller, "Cyber Security Capacity Building in Developing Countries: Challenges and Opportunities," Norwegian Institute of International Affairs, Oslo, Norway, 3, 2015. [Online]. Available: https://cybilportal.org/wp-content/uploads/2020/06/NUPIReport03-15-Muller.pdf. [Accessed: Nov. 13, 2023].

[10] S. Creese, W. H. Dutton, P. Esteve-González, "The social and cultural shaping of cybersecurity capacity building: a comparative study of nations and regions," *Personal and Ubiquitous Computing*, vol. 25, pp. 941–955, 2021, doi: 10.1007/s00779-021-01569-6.

[11] Z. Homburger, "The Necessity and Pitfall of Cybersecurity Capacity Building for Norm Development in Cyberspace," *Global Society*, vol. 33, no. 2, pp. 224–242, 2019, doi: 10.1080/13600826.2019.1569502.

[12] S. Almuhammadi, M. Alsaleh, "Information Security Maturity Model for Nist Cyber Security Framework," *Computer Science & Information Technology* (CS & IT), Academy & Industry Research Collaboration Center (AIRCC), pp. 51-62, 2017, doi: 10.5121/csit.2017.70305.

[13] Global Cyber Security Capacity Centre, "Cybersecurity Capacity Maturity Model for Nations (CMM)," 2021. [Online]. Available: https://gcscc.ox.ac.uk/the-cmm. [Accessed: April 11, 2023].

[14] Freedom House, "Freedom in the World 2023," 2023. [Online]. Available: https://freedomhouse.org/report/freedom-world/2023/marking-50-years. [Accessed: Nov. 13, 2023].

[15] D. Kaufmann, A. Kraay, M. Mastruzzi, "The Worldwide Governance Indicators: Methodology and Analytical Issues," *Hague Journal of the Rule of Law*, vol. 3, no. 2, pp. 220–246, 2011, doi: 10.1017/S1876404511200046.

[16] J. Voo, I. Hemani, D. Cassidy, *National Cyber Power Index 2022*. Cambridge, MA: Belfer Center for Science and International Affairs, Harvard Kennedy School, 2022.

[17] A. de Tocqueville, *Democracy in America.* Washington, D.C.: Regnery Publishing, 2003.

[18] G. A. Almond, S. Verba, *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Princeton, NJ: Princeton University Press, 2015.

[19] OECD, "Recommendations of the Council on Digital Security Risk Management," Paris, 2022. [Online]. Available: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0479. [Accessed: Apr. 12, 2023].

[20] National Cyber Security Centre, "NCSC CAF Guidance Principles and Related Guidelines," 2019. [Online]. Available: https://www.ncsc.gov.uk/collection/caf/table-view-principles-and-related-guidance. [Accessed: Apr. 12, 2023].

[21] NIST, "Framework for Improving Critical Infrastructure Cybersecurity," National Institute of Standards and Technology, pp. 1–41, 2014, doi: 10.6028/NIST.CSWP.04162018.

[22] K. L. Kraemer, W. H. Dutton, A. Northrop, *The Management of Information Systems*. New York: Columbia University Press, 1981, doi: 10.7312/krae93774.

[23] K. L. Kraemer, J. N. Danziger, W. H. Dutton, A. M. Mood, R. Kling, "A future cities survey research design for policy analysis," *Socio-Economic Planning Science*, vol. 10, no. 5, pp. 199–211, 1976, doi: 10.1016/0038-0121(76)90029-X.

[24] J. N. Danziger, W.H. Dutton, R. Kling, K. L. Kraemer, *Computers and politics. High technology in American local governments*. New York: Columbia University Press, 1982.

[25] A. Ryan, *On Tocqueville: Democracy and America*. New York: W. W. Norton & Company, 2014.

[26] J. Corbin, A. Strauss, "Grounded Theory Research : Procedures , Canons and Evaluative Criteria," *Zeitschrift Für Soziologie.*, vol. 19, no. 6, pp. 418–427, 1990, doi: 10.1515/zfsoz-1990-0602.

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Predictive Modelling of a Honeypot System Based on a Markov Decision Process and a Partially Observable Markov Decision Process

Lidong Wang | Institute for Systems Engineering Research Mississippi State University, Mississippi, USA, ORCID: 0000-0003-3923-849X
Reed Mosher | Institute for Systems Engineering Research Mississippi State University, Mississippi, USA
Patti Duett | Institute for Systems Engineering Research Mississippi State University, Mississippi, USA
Terril Falls | Institute for Systems Engineering Research Mississippi State University, Mississippi, USA, ORCID: 0009-0006-4468-4928

Abstract

A honeypot is used to attract and monitor attacker activities and capture valuable information that can be used to help practice good cybersecurity. Predictive modelling of a honeypot system based on a Markov decision process (MDP) and a partially observable Markov decision process (POMDP) is performed in this paper. Analyses over a finite planning horizon and an infinite planning horizon for a discounted MDP are respectively conducted. Four methods, including value iteration (VI), policy iteration (PI), linear programming (LP), and Q-learning, are used in the analyses over an infinite planning horizon for the discounted MDP. The results of the various methods are compared to evaluate the validity of the created MDP model and the parameters in the model. The optimal policy to maximise the total expected reward of the states of the honeypot system is achieved, based on the MDP model employed. In the modelling over an infinite planning horizon for the discounted POMDP of the honeypot system, the effects of the observation probability of receiving commands,

Received: 27.11.2022

Accepted: 02.01.2023

Published: 04.01.2023

Cite this article as:
L. Wang, R. Mosher, P. Duett, T. Falls "Predictive Modelling of a Honeypot System Based on a Markov Decision Process and a Partially Observable Markov Decision Process," ACIG, vol. 2, no. 1, 2023, DOI: 10.5604/01.3001.0016.2027

Corresponding author:
Lidong Wang, Institute for Systems Engineering Research Mississippi State University, Mississippi, USA; ORCID: 0000-0003-3923-849X; E-MAIL: lidong@iser.msstate.edu

the probability of attacking the honeypot, the probability of the honeypot being disclosed, and transition rewards on the total expected reward of the honeypot system are studied.

Keywords
cybersecurity, honeypot, machine learning, Markov decision process, partially observable Markov decision process, Q-learning

1. Introduction

Cybersecurity is concerned with the privacy and security of computers or electronic devices, networks, and any information that is stored, processed, or exchanged by information systems [1]. Parameter design, monitoring, and network maintenance are important to network cybersecurity. The detection and prevention of attacks are generally more significant than any subsequent actions taken after being attacked [2]. It is helpful to obtain as much information as possible from attacks to defend against attackers and improve the cybersecurity of information systems [3]. A honeypot system can collect information from an attack about the attackers and may aid in the practice of robust cybersecurity. A honeypot is used to attract attackers and record their activities [4].

Attackers can be attracted to a fake system by a honeypot in the network infrastructure; valuable information can be obtained from them; and the information can then be used to improve network security [4]. A honeypot constitutes a useful technique or tool to observe the spread of malware and the emergence of new exploits. An attacker tries to avoid connecting to a honeypot as it can disclose the attacker's tools, methods, and exploits [5]. A honeypot is also a source that can be leveraged to build high-quality intelligence against threats, providing a means for monitoring attacks and discovering zero-day exploits [6]. A network honeypot is often used by information security teams to measure the threat landscape for the security of their networks [7]. One example of a stochastic process method, the MDP, has been used for decision-making in cybersecurity. The MDP assumes that both defenders and attackers have observable information, although this is not true in many applications [8]. In actuality, there may be partial observability or an agent's inability to fully observe the state of its environment in numerous real situations [9]. In many real-world problems, their environmental models are not known. There is a considerable need for reinforcement learning to solve problems where agents partially observe the states of their environments (possibly due to noise in the observed data). This leaves the outcomes of actions under uncertainty more dependent on the signal of the current

state. The POMDP extends the MDP by permitting a decision-making process under uncertain or partial observability [10]. The artificial intelligence (AI) world has shown a huge leap recently in the research area of the POMDP model [11].

An MDP model for interaction honeypots was created and an analytic formula of the gain was derived. The optimal policy was decided based on comparing the calculated gain of each policy and selecting the one with a maximal gain. The model was then extended using a POMDP. One approach to solving the POMDP problem was proposed. In this method, the system state was replaced with the belief state and the POMDP problem was converted into an MDP problem [12]. The efforts in the research of this paper were to fulfil predictive modelling of the honeypot system, based on the MDP and the POMDP. Various methods and algorithms were used, including VI, PI, LP, and Q-learning in the analyses of the discounted MDP over an infinite planning horizon. The results of these algorithms were evaluated to validate the created MDP model and its parameters. In the modelling of the discounted POMDP over an infinite planning horizon, the effects of several important parameters on the system's total expected reward were studied. These parameters include the observation probability of receiving commands, the probability of attacking the honeypot, the probability of the honeypot being disclosed, and the transition rewards. The analyses of the MDP and POMDP in this paper were conducted using the R language and R functions. This paper is organised as follows: the second section introduces the methods of MDP and POMDP; Section 3 presents a created MDP model of the system and the parameters in the model; Section 4 shows the analyses of the system based on the MDP method; Section 5 presents analyses of the system based on the POMDP method, and the final section is the conclusion.

2. Methods

2.1 The MDP

The MDP method is one of the most significant methods employed in artificial intelligence (especially machine learning). The MDP is described using the tuple $<S, A, T, R, \gamma>$ [13–15]:

- S is the states' set.
- A is the actions set.
- T is the transition probability from the state s to the state s' ($s \in S$, $s' \in S$) after action a ($a \in A$).
- R is an immediate reward after action a, and
- γ ($0 < \gamma < 1$) is the discounted factor.

An optimal policy is the goal of the MDP that maximises the total expected reward. An optimal policy over a finite planning horizon maximises the vector of the total expected reward until the horizon ends. The total expected reward (discounted) for an infinite planning horizon is employed to evaluate the gain of the discounted MDP in this paper.

2.2. The Algorithms of the MDP

VI, PI, LP, and Q-learning have been the algorithms utilised to find an optimal policy for the MDP. Theoretically, the results of the four kinds of algorithms should be the same. However, the results obtained using the algorithms may potentially differ with a great value, or convergence problems may potentially occur during the iterative process if the created MDP model is unreasonable, owing to unsuitable structure or incorrect model parameters. Thus, all the algorithms are employed, and their results are evaluated to validate the model constructed in this paper.

VI: An optimal policy for the MDP can be achieved by utilizing VI when the planning horizon is finite. In principle, the four algorithms (VI, PI, LP, and Q-learning) can be employed to find the optimal policy when the planning horizon is infinite. VI utilises the following equation of value iterations [16–18] to calculate the total expected reward for each state:

$$V(s) := max_a \sum_{s'} T(s, a, s')(R(s, a, s') + \gamma V(s')) \quad (1)$$

where $T(s, a, s')$ is the transition probability from state s to state s' after action a. $R(s, a, s')$ is the immediate reward of the transition. $V(s)$ and $V(s')$ are the total expected reward in state s and state s', respectively. When the value difference between two consecutive iterative steps is lower than the given tolerance, the iteration will be stopped.

PI: A better policy is found using PI, through comparing the current policy to the previous one. PI generally begins arbitrarily with an initial policy and then policy evaluation and policy improvement are followed. The process of iterations continues until the same policy is obtained for two successive policy iterations, indicating that the optimal policy has been achieved. For each state s, Equation (2) is used for policy evaluation and Equation (3) is used for updating the policy (policy improvement) [16, 18].

$$V(s) := max_a \sum_{s'} T(s, \pi(s), s')(R(s, \pi(s), s') + \gamma V(s')) \quad (2)$$

where $\pi(s)$ is an optimal policy of state s.

$$\pi(s) = argmax_a \left(\Sigma_{s'} T(s, a, s')(R(s, a, s') + \gamma V(s'))\right) \quad (3)$$

LP: Since the MDP can be expressed as a linear program, the LP can find a static policy through solving the linear program. The following LP formulation [19] is used to find the optimal value function:

Solve

$$\min_V \Sigma_{s \forall S} V(s) \quad (4)$$

subject to

$$V(s) \geq R(s, a, s') + \gamma \Sigma_{s' \in S} T(s, a, s') V(s') \quad (5)$$

Q-LEARNING: It is used to achieve the best policy with the greatest reward. It is a reinforcement learning method and allows an agent to learn the Q-value function that is an optimal action-value function. Q-learning can also be applied to non-MDP domains [20]. The action-value function

$Q(s, a)$ is expressed as follows [21]:

$$Q(s, a) = \Sigma_{s'} T(s, a, s')(R(s, a, s') + \gamma V(s')) \quad (6)$$

$Q(s, a)$ can be initialised arbitrarily (for example, $Q(s, a) = 0$, $\forall s \in S$, $\forall a \in A$). From state s to state s', a Q-learning update can be defined as follows [21, 22]:

$$Q(s, a) := (1 - \beta)Q(s, a) + \beta [R(s, a) + \gamma \max_a Q(s', a)] \quad (7)$$

where $\beta \in (0, 1)$ represents the learning rate. The best action a at state s can be chosen according to the optimal policy $\pi(s)$. The iterative process continues until the final step of episode. The optimal policy is described as follows:

$$\pi(s) = \arg \max_{a \in A} Q(s, a) \quad (8)$$

2.3. The POMDP

A POMDP can be thought as a generalisation of an MDP, permitting state uncertainty in a Markov process [23]. In POMDP applications, the objective is generally to obtain a decision rule or policy to maximise the expected long-term reward [24]. In the POMDP, the belief state is a distribution of probabilities over all possible states. An optimal action relies only on the current belief state [25].

The POMDP was defined as a tuple <S, A, T, R, O, B, γ> [26]:

- O = {o_1, o_2, ...,o_k} is an observation set.
- B is a set of conditional observation probabilities $B(o|s', a)$. s' is the new state after the state transition $s \rightarrow s'$, $o \in O$.
- S, A, T, R, and γ are the same as those in the tuple of MDP.

After having taken the action a and observing o, the belief state needs to be updated. If $b(s)$ is the previous belief state, then the new belief state [25]) is given by

$$b'(s') = \alpha P(o|s') \sum_s P(s'|s, a)\, b(s) \quad (9)$$

where α is a normalizing constant that makes the belief state sum to 1.

The goal of POMDP planning is to obtain a sequence of actions {a_0, a_2, ...,a_t} at time steps that maximise the total expected reward [27], i.e., we choose actions that give

$$\max E\,[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)] \quad (10)$$

where s_t and a_t are the state and the action at time t, respectively.

The optimal policy brings up the greatest expected reward for each belief state, which is the solution to the Bellman optimality equation through iterations beginning at an initial value function for an initial belief state. The equation can be formulated as [12]:

$$V(b) = \max_{a \in A} [b(s)R(s, a) + \gamma \sum_{o \in O} P(o|b,a)V(b')] \quad (11)$$

3. The MDP Model of the Honeypot System

3.1 The Structure of the MDP Model

The honeypot system is a network-attached system that is put in place to lure attackers. A botnet is utilised to forward spam, steal data, etc. A botmaster keeps a bot online. A honeypot has three states [12]:

- State 1: Not attacked yet (waiting for an attack to join the botnet).
- State 2: Compromised (becoming a member of the botnet).
- State 3: Disclosed (not the botnet's member anymore) due to the real identity having been discovered or interactions

with the botmaster having been lost for an extended period of time.

A honeypot can take one of the following actions at each state:

- Action 1: Allows a botmaster to compromise the honeypot system and to implement commands.
- Action 2: Does not allow the botmaster to compromise the system.
- Action 3: Reinitialised as a new honeypot and reset to the initial state.

A model of the honeypot system is established based on the MDP. Fig. 1 shows the state transitions of the states (1, 2, and 3) resulted from each of the actions (Action 1, Action 2, and Action 3).

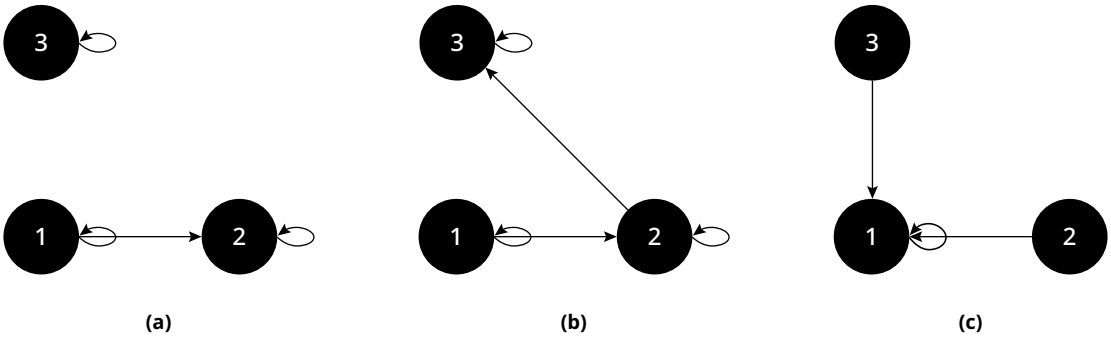

Figure 1. The state transitions due to each of the three actions: (a) Action 1, (b) Action 2, and (c) Action 3.

3.2. State Transition Matrix and Reward Matrix

The transitions between the states in the created model of the system rely on one of the actions and on two important probabilities [12]. State 1 cannot be transitioned to State 3 directly; State 3 cannot be transitioned to State 2. The probability of a transition from State 3 to State 1 is 0 (under Action 1 or Action 2) or 1 (under Action 3). The following is a description of the two important probabilities:

1. P_a: the probability of attacking the honeypot.
2. P_d: the probability of the honeypot being disclosed.

The benefit and expenses due to the state transitions or self-transitions are as follows [12]:

1. E_o: the operation expense due to running, deploying, and controlling a honeypot.

2. E_r: the expense in reinitializing a honeypot.

3. E_l: the expense in liability when a honeypot operator becomes liable for implementing a botmaster's commands if those commands include illicit actions.

4. B_i: the benefit of information when a honeypot collects an attacker's information regarding techniques, codes, and tools.

The state transition probability matrix T and the reward matrix R under each action are formulated as follows:

1. T and R under Action 1 are

$$T = \begin{bmatrix} 1-P_a & P_a & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{12}$$

$$R = \begin{bmatrix} -E_0 & B_i - E_0 & 0 \\ 0 & B_i - E_0 - E_l & 0 \\ 0 & 0 & -E_0 \end{bmatrix} \tag{13}$$

2. T and R under Action 2 are

$$T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1-P_d & P_d \\ 0 & 0 & 1 \end{bmatrix} \tag{14}$$

$$R = \begin{bmatrix} -E_0 & 0 & 0 \\ 0 & B_i - E_0 & B_i - E_0 \\ 0 & 0 & -E_0 \end{bmatrix} \tag{15}$$

3. T and R under Action 3 are

$$T = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{16}$$

$$R = \begin{bmatrix} -E_r & 0 & 0 \\ B_i - E_r & 0 & 0 \\ -E_r & 0 & 0 \end{bmatrix} \tag{17}$$

4. Analyses of the Honeypot System Based on MDP

4.1 MDP-based Analyses over an Infinite Planning Horizon

Let P_a= 0.6, P_d= 0.6, E_o= 1, E_r= 2.5, B_i= 16, E_l= 14, and γ = 0.85. Analyses are performed using the R language and its functions. By substituting the data into equations (12–17), the values of T and R under various actions (due to various policies) can be computed:

T and R under Action 1 become

$$T = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} -1 & 15 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

T and R under Action 2 are

$$T = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0.4 & 0.6 \\ 0 & 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 15 & -17 \\ 0 & 0 & -1 \end{bmatrix}$$

T and R under Action 3 are

$$T = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} -2.5 & 0 & 0 \\ -18.5 & 0 & 0 \\ -2.5 & 0 & 0 \end{bmatrix}$$

Various policies are evaluated, and Tab. 1 shows the result of the total expected rewards for states with various policies. For example, the policy c (1, 1, 3) indicates that Action 1, Action 1, and Action 3 are taken on State 1, State 2, and State 3, respectively. $V1$, $V2$, and $V3$ represent the total expected reward for State 1, State 2, and State 3, respectively.

Table 1. The total expected reward of each state for four various policies (γ = 0.85).

Policy	c (1, 1, 2)	c (1, 1, 3)	c (1, 2, 3)	c (2, 1, 3)
V1	18.1818	18.1818	13.4431	-6.6667
V2	6.6667	6.6667	0.5342	6.6667
V3	-6.6667	12.9545	8.9266	-8.1667

The four kinds of algorithms (VI, PI, LP, and Q-learning) can be implemented using the values of T and R under various actions. These algorithms are used in this paper and the optimal policy achieved

using the four algorithms is c (1, 1, 3) in each case. The results for the total expected rewards for each state are compared to evaluate the validity of the MDP model in this paper. The results of the honeypot system (based on a discounted MDP with γ = 0.85) over an infinite planning horizon are shown in Tab. 2.

VI consists of solving Bellman's equation iteratively. Jacob's algorithm and Gauss-Seidel's algorithm are employed in the VI method respectively, so that there are two variants of VI algorithm employed. In Gauss-Seidel's value iterations, $V(k+1)$ is used instead of $V(k)$ whenever this value has been calculated; k is the iteration number. In this situation, the convergence speed is enhanced. It is also shown that its accuracy is improved in comparison to Jacob's algorithm (Tab. 2.). The result of Gauss-Seidel's value iteration algorithm shows that the total expected reward is 18.1818 (the highest value) if the MDP starts in state 1 while it is 6.6667 (the lowest value) if the MDP starts in state 2. The Q-learning result in Table 2 was obtained when the number of iterations was 150,000. The results of the VI (Gauss-Seidel algorithm), PI, and LP are the same, and very close to the Q-learning result, indicating the MDP model created is valid, and that the model parameters are indeed suitable.

Table 2. Analyses of the honeypot system based on various algorithms over an infinite planning horizon (γ = 0.85)

Algorithm	*V1*	*V2*	*V3*
VI (Jacob algorithm)	17.9622	6.4470	12.7349
VI (Gauss-Seidel algorithm)	18.1818	6.6667	12.9545
PI	18.1818	6.6667	12.9545
LP	18.1818	6.6667	12.9545
Q-learning	18.1699	6.6667	12.9206

4.2. The MDP-based Analysis for the Honeypot System over a Finite Planning Horizon

The above data regarding probabilities, the benefit, and expenses (i.e., P_a, P_d, E_o, E_r, B_i, and E_l) are also utilised in the analysis of the system with the discount γ = 0.85 over a finite planning horizon based on the MDP method. Tab. 3 shows the total expected rewards of the three states that were calculated using value iterations over a 50-step planning horizon. $V1(n)$, $V2(n)$, and $V3(n)$ are the total expected reward at step n for State 1, State 2, and State 3, respectively.

It is shown that the total expected rewards $V1(n)$, $V2(n)$, and $V3(n)$ are very close to $V1$, $V2$, and $V3$ for the infinite planning horizon in Tab. 2 when epoch $n \leq 20$.

Table 3. Total expected rewards for three states calculated using value iterations over a 50-step planning horizon ($\gamma = 0.85$).

Epoch n	$V1(n)$	$V2(n)$	$V3(n)$
0	18.1798	6.6647	6.6647
5	18.1774	6.6622	6.6622
10	18.1718	6.6567	6.6567
15	18.1592	6.6441	6.6441
20	18.1309	6.6158	6.6158
25	18.0672	6.5520	6.5520
30	17.9234	6.4083	6.4083
35	17.5995	6.0843	6.0843
40	16.8691	5.3542	5.3542
45	15.1715	3.7086	3.7086
46	14.5479	3.1866	3.1866
47	13.6351	2.5725	2.5725
48	12.0340	1.8500	1.8500
49	8.6	1.0	1.0
50	0	0	0

5. Analyses of the Honeypot System Based on the POMDP

5.1 Observations and Observation Probabilities in the Honeypot System

The POMDP model of the system is based on the MDP model shown in Fig. 1, and observations as well as observation probabilities are considered to model uncertainty in the POMDP model. Three observations [12] are employed to compute and monitor the system belief state:

- *Unchanged*: The honeypot does not have any observed change, indicating it is still in the waiting state (State 1).

- *Absence*: It means an absence of botmasters' commands after the honeypot was compromised. This situation can be due to 1) the honeypot being detected and disconnected from the botnet, or 2) botmasters being busy with other things (for example, compromising other machines), leading to uncertainty in determining whether the honeypot is in State 2 (compromised) or State 3 (disclosed).

- *Commands*: After the honeypot is compromised, it receives the command information from a botmaster, indicating that it is not disclosed yet and still in State 2.

In State 2, the probability of receiving commands is denoted by $Po1$, while the probability of absence is denoted by $Po2$. Therefore, we have the following observation probabilities:

For the honeypot in State 1:
$P(Unchanged) = 1$, $P(Commands) = P(Absence) = 0$

For the honeypot in State 2:
$P(Unchanged) = 0$, $P(Commands) = Po1$
$P(Absence) = Po2 = 1 - Po1$

For the honeypot in State 3:
$P(Unchanged) = P(Commands) = 0$, $P(Absence) = 1$

5.2. Analyses Based on Various Solution Methods of the POMDP over An Infinite Planning Horizon

Analyses over an infinite planning horizon for a discounted POMDP of the honeypot system are performed. Let $P_a = 0.6$, $P_d = 0.6$, $E_o = 1$, $E_r = 2.5$, $B_i = 16$, $E_l = 14$, and $\gamma = 0.85$. The following solution methods or algorithms [23, 24, 26–29, 30] are used to solve the POMDP problem: Grid, Enumeration, Two Pass, Witness, Incremental Pruning, and SARSOP. The total expected reward of the honeypot system based on POMDP is denoted by V_t in this paper. The values of V_t at three different observation probabilities of receiving commands ($Po1 = 0.5$, 0.6, and 0.7) are computed using various solution methods of POMDP. The result of V_t is shown in Tab. 4. The values of Incremental Pruning and SARSOP are very close to the results of the other four methods and the results of the four methods are the same.

Table 4. The total expected reward of the honeypot system based on various solution methods of POMDP.

Methods	V_t (P_{o1} = 0.5)	V_t (P_{o1} = 0.6)	V_t (P_{o1} = 0.7)
Grid	9.850447	10.187263	10.449232
Enumeration	9.850447	10.187263	10.449232
Two Pass	9.850447	10.187263	10.449232
Witness	9.850447	10.187263	10.449232
Incremental Pruning	9.848475	10.185292	10.447260
SARSOP	9.850403	10.187213	10.449210

5.3. The Analysis for the Honeypot System with Various Observation Probabilities of Receiving Commands

The total expected reward V_t of the honeypot system with various observation probabilities of receiving commands ($Po1$) is analysed for the discounted POMDP over an infinite planning horizon. Grid is used to solve the POMDP problem. It tries to approximate the value function over an entire state space according to the estimation for a finite number of belief states on the chosen grid [31]. The following data are used in the analysis: P_a= 0.6, P_d= 0.6, E_o= 1, E_r= 2.5, B_i= 16, E_l= 14, and γ = 0.85; $Po1$ = 0.1, 0.2, 0.3, ..., 0.9. Figure 2 shows that the total expected reward V_t of the honeypot system increases as the observation probability ($Po1$) of receiving commands rises. In the following sections of this paper, the Grid method is also used in solving the POMDP problem.

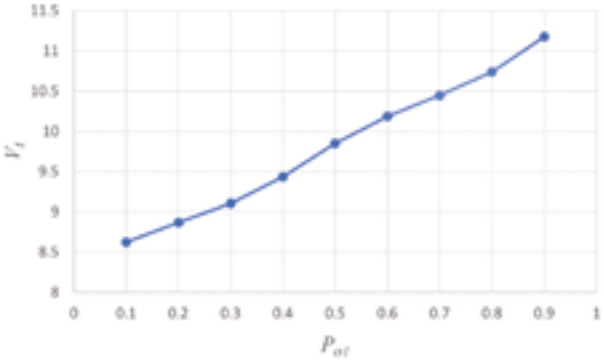

Figure 2. The total expected reward Vt of the honeypot system at various $Po1$.

5.4. Analyses for the System with Various P_a and P_d

An analysis for the discounted POMDP with various P_a over an infinite planning horizon is conducted. The following data are utilised:

P_d= 0.6, E_o= 1, E_r= 2.5, B_i= 16, E_l= 14, and γ = 0.8 5. The total expected reward V_t of the honeypot system at various P_a for various $Po1$ is analysed and the result is shown in Fig. 3. V_t increases with higher values of P_a, although the rate of increase steadily diminishes. The increased P_a provides the honeypot with more opportunities for collecting valuable information about attackers. V_t is larger when $Po1$ is larger.

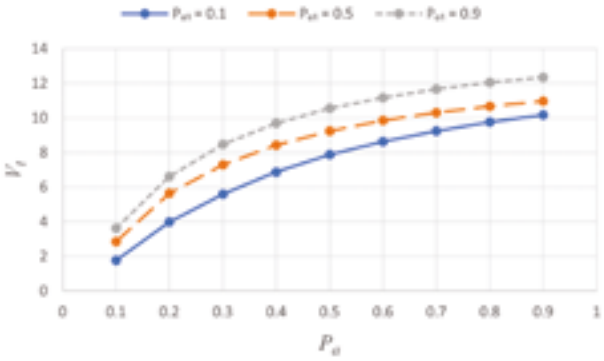

Figure 3. The total expected reward V_t of the honeypot system at various P_a.

Let P_a= 0.6, E_o= 1, E_r= 2.5, B_i= 16, E_l= 14, and γ = 0.85. The V_t at various P_d for various $Po1$ is analysed over an infinite planning horizon, and Figure 4 shows the results. V_t is higher when $Po1$ is higher, but the value of V_t when $Po1$ = 0.1 is very close to that of V_t when $Po1$ = 0.5 (if P_d < 0.5). For $Po1$ = 0.1, V_t falls as P_d is increased from 0.1 to 0.8 and is unchanged when P_d moves from 0.8 to 0.9; for P_{oc} = 0.5, V_t decreases as P_d is increased from 0.1 to 0.6 and is unchanged as P_d goes from 0.6 to 0.9; for $Po1$ = 0.9, V_t declines as P_d is increased from 0.1 to 0.5, though it does not change as P_d moves from 0.5 to 0.9. There is no significant difference in V_t for $Po1$ = 0.5 and $Po1$ = 0.9 when P_d changes from 0.5 to 0.9.

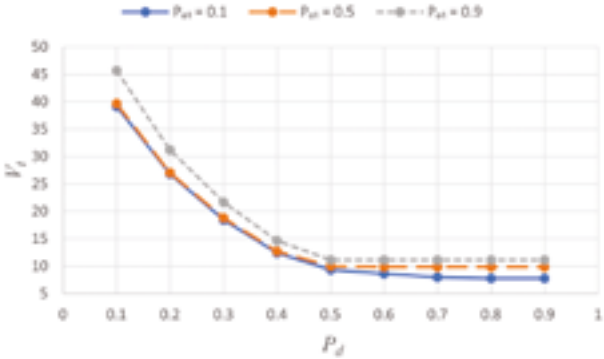

Figure 4. The total expected reward V_t of the honeypot system at various P_d.

5.5. Analyses for the System with Various Transition Rewards

Analyses for the honeypot system with various transition rewards over an infinite planning horizon are performed. The following data are utilised: P_a= 0.6, P_d= 0.6, E_o= 1, E_r= 2.5, E_l= 14, and γ = 0.85. The total expected reward V_t at various P_a for various $Po1$ is analysed, and the results are shown in Fig. 5. V_t initially increases slightly (B_i < 14) and then more rapidly (B_i > 14) with the increase of B_i. V_t for various $Po1$(0.1, 0.5, and 0.9) is the same when B_i = 10, 11, and 12. V_t is the same for $Po1$ = 0.1 and 0.5 when B_i = 13. When Bi > 13, V_t is larger if P_{oc} is larger.

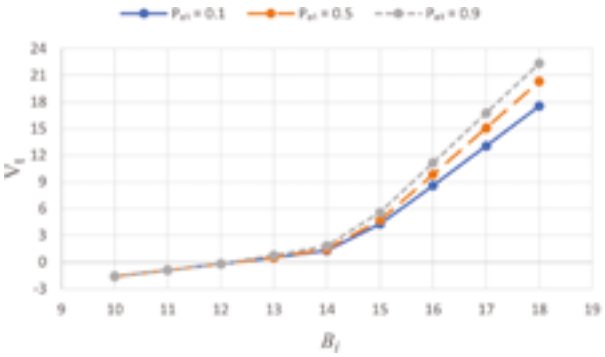

Figure 5. The total expected reward V_t of the honeypot system V_t at various B_i.

Let P_a= 0.6, P_d= 0.6, E_o= 1, E_r= 2.5, B_i= 16, and γ = 0.85. The total expected reward V_t at various E_l for various $Po1$ is analysed over an infinite planning horizon and Figure 6 shows the results. V_t decreases when E_l is increased from 12 to 16. V_t is the same for $Po1$ = 0.1 and 0.5 as E_l rises from 17 to 20. It is the same for all the three values of $Po1$(0.1, 0.5, and 0.9) when E_l goes from 19 to 20.

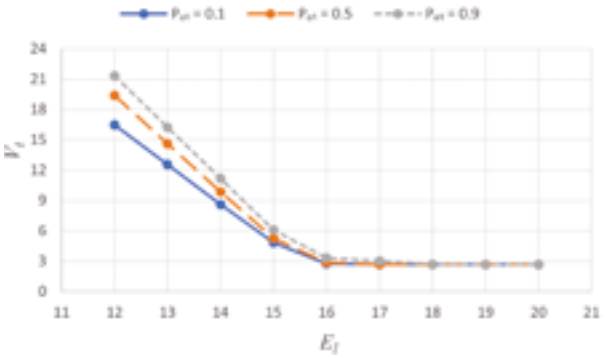

Figure 6. The total expected reward V_t of the honeypot system at various E_l.

6. Conclusion

The MDP-based predictive modelling for the honeypot system has demonstrated that the model and algorithms in this paper are suitable for performing analyses over both a finite planning horizon and an infinite planning horizon (for a discounted MDP), and that they are effective at finding an optimal policy and maximizing the total expected rewards of the states of the honeypot system. The results of the total expected reward using Gauss-Seidel's algorithm of VI, PI, and LP are the same, and the result of Q-learning is very close to the same result, indicating the MDP model created in this paper is valid and that the model parameters are suitable.

In the predictive modelling of the honeypot system based on the discounted POMDP over an infinite planning horizon, the total expected reward V_t of the honeypot system increases with the increase of the observation probability of receiving commands ($Po1$). It also rises as P_a is increased or B_i is increased. The increased P_a leads to more opportunities for the honeypot to collect valuable information about attackers. As P_d increases, V_t declines at first and then levels out. As E_l increases, V_t decreases by successively smaller amounts until it eventually flattens out.

Declaration of Competing Interest

The authors in this paper do not have any competing interest.

Data availability

No database or dataset was used or generated in the research of this article.

Acknowledgments

This paper is based upon work supported by Mississippi State University (MSU) in the USA.

References

[1] McKenzie, T. M. (2017). *Is Cyber Deterrence Possible?, Air University Press*.

[2] S. Srujana, P. Sreeja, G. Swetha, H. Shanmugasundaram, "Cutting edge technologies for improved cybersecurity model: A survey," *International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2022, pp. 1392–1396.

[3] L. Vokorokos, A. Pekár, N. Ádám, P. Daránvi, "Yet another attempt in user authentication," *Acta Polytechnica Hungarica*, vol. 10, no. 3, pp. 37–50, 2013.

[4] J. Palša, J. Hurtuk, E. Chovancová, M. Havira, "Configuration honeypots with an emphasis on logging of the attacks and redundancy," *IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2022, pp. 000073–000076, doi: 10.1109/SAMI54271.2022.9780801

[5] F. Franzen, L. Steger, J. Zirngibl, P. Sattler, "Looking for honey once again: Detecting RDP and SMB honeypots on the Internet," *IEEE Looking European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2022.

[6] M. Boffa, G. Milan, L. Vassio, I. Drago, M. Mellia, Z.B. Houidi, "Towards NLP-based processing of honeypot logs," *IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2022, pp. 314–321

[7] Z. Shamsi, D. Zhang, D. Kyoung, A. Liu, „Measuring and Clustering Network Attackers using Medium-Interaction Honeypots," in *IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2022, pp. 294–306.

[8] X. Liu, H. Zhang, S. Dong, Y. Zhang, „*Network Defense Decision-Making Based on a Stochastic Game System and a Deep Recurrent Q-Network,*" *Computers & Security*, vol. 111, p. 102480, 2021, doi: 10.1016/j.cose.2021.102480.

[9] H. Itoh, H. Nakano, R. Tokushima, H. Fukumoto, „A Partially Observable Markov Decision Process-Based Blackboard Architecture for Cognitive Agents in Partially Observable Environments," *IEEE Transactions on Cognitive and Developmental Systems*, 2020, doi: 10.1109/TCDS.2020.3034428.

[10] M. Haklidir, H. Temeltaş, „Guided Soft Actor Critic: A Guided Deep Reinforcement Learning Approach for Partially Observable Markov Decision Processes," *IEEE Access*, vol. 9, pp. 159672–159683, 2021, doi: 10.1109/ACCESS.2021.3131772.

[11] A.R. Cassandra, „A Survey of POMDP Applications," 2003. [Online]. Available: http://www.cassandra.org/arc/papers/applications.pdf. [Accessed: Nov. 27, 2022].

[12] O. Hayatle, H. Otrok, A. Youssef, „A Markov Decision Process Model for High Interaction Honeypots," *Information Security Journal: A Global Perspective*, vol. 22, no. 4, pp. 159–170, 2013.

[13] M. Mohri, A. Rostamrdeh, A. Talwalkar, *Foundations of Machine Learning*. Cambridge, Massachusetts: MIT Press, 2012.

[14] M.A. Alsheikh, D.T. Hoang, D. Niyato, H.P. Tan, S. Lin, „Markov Decision Processes with Applications in Wireless Sensor Networks: A Survey," *IEEE Communications*

Surveys & Tutorials, vol. 17, no. 3, pp. 1239–1267, 2015, doi: 10.1109/COMST.2015.2420686.

[15] Y. Chen, J. Hong, C.C. Liu, „Modeling of Intrusion and Defense for Assessment of Cyber Security at Power Substations," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2541–2552, 2018.

[16] R.S. Sutton, A.G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, Massachusetts: MIT Press, 2018.

[17] M. van Otterlo, „Markov Decision Processes: Concepts and Algorithms," in *Reinforcement Learning: Adaptation, Learning, and Optimization*, vol. 12, M. Wiering and M. van Otterlo, Eds. Berlin, Heidelberg: Springer, doi: 10.1007/978-3-642-27645-3_1.

[18] M. van Otterlo, M. Wiering, „Reinforcement Learning and Markov Decision Processes," in *Reinforcement Learning: Adaptation, Learning, and Optimization*, vol. 12, M. Wiering and M. van Otterlo, Eds. Berlin, Heidelberg: Springer, pp. 3–42, doi: 10.1007/978-3-642-27645-3_1.

[19] O. Sigaud, O. Buffet, *Markov Decision Processes in Artificial Intelligence*. Hoboken, New Jersey: John Wiley & Sons, 2013.

[20] S.J. Majeed, M. Hutter, „On Q-learning Convergence for Non-Markov Decision Processes," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2546–2552, 2018, doi: 10.24963/ijcai.2018/353.

[21] E. Zanini, *Markov Decision Processes*. Berlin, Heidelberg: Springer, 2014.

[22] Y. Liu, H. Liu, B. Wang, „Autonomous Exploration for Mobile Robot Using Q-Learning," in *Proceedings of the 2nd International Conference on Advanced Robotics and Mechatronics (ICARM)*, 2017, pp. 614–619, doi: 10.1109/ICARM.2017.8273233.

[23] G.E. Monahan, „State of the Art–A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms," *Management Science*, vol. 28, no. 1, pp. 1–16, 1982, doi: 10.1287/mnsc.28.1.1.

[24] M.L. Littman, A.R. Cassandra, L.P. Kaelbling, „Efficient Dynamic-Programming Updates in Partially Observable Markov Decision Processes," Department of Computer Science. Providence, Rhode Island: Brown University, CS-95-19, 1995.

[25] J. Stuart, P. Norvig, *Artificial Intelligence: A Modern Approac*h, London: Pearson, 3rd ed., 2010.

[26] H. Kurniawati, D. Hsu, W.S. Lee, „Sarsop: Efficient Point-Based POMDP Planning by Approximating Optimally Reachable Belief Spaces," in *Robotics: Science and Systems*, 2008.

[27] J. Pineau, G. Gordon, S. Thrun, „Point-Based Value Iteration: An Anytime Algorithm for POMDPs," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 3, pp. 1025–1032, 2003.

[28] E.J. Sondik, *The Optimal Control of Partially Observable Markov Processes*. Stanford, California: Stanford University, 1971.

[29] N.L. Zhang, W. Liu, „Planning in Stochastic Domains: Problem Characteristics and Approximation," Department of Computer Science. Hong Kong: Hong Kong University of Science and Technology, HKUST-CS96-31, 1996.

[30] A.R. Cassandra, M.L. Littman, N.L. Zhang, „Incremental Pruning: A Simple, Fast, Exact Method for Partially Observable Markov Decision Processes," *arXiv preprint arXiv:1302.1525*, 2013.

[31] R.I. Brafman, „A Heuristic Variable Grid Solution Method for POMDPs," in *Proceedings of the AAAI/IAAI*, 1997, pp. 727–733.

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Artificial Immune Systems in Local and Network Cybersecurity: An Overview of Intrusion Detection Strategies

Patryk Widuliński | Faculty of Electronics and Computer Science, Koszalin University of Technology, Poland, ORCID: 0000-0001-7258-3522

Received: 24.10.2023

Accepted: 30.11.2023

Published: 07.12.2023

Cite this article as:
P. Widuliński "Artificial Immune Systems in Local and Network Cybersecurity: An Overview of Intrusion Detection Strategies," ACIG, vol. 2, no. 1, 2023. DOI: 10.60097/ACIG/162896.

Corresponding author:
Patryk Widuliński, Faculty of Electronics and Computer Science, Koszalin University of Technology, Poland; ORCID: 0000-0001-7258-3522; E-MAIL: patryk.widulinski@tu.koszalin.pl

Abstract

In this paper, an overview of artificial immune systems (AIS) used in intrusion detection systems (IDS) is provided, along with a review of recent efforts in this field of cybersecurity. In particular, the focus is on the negative selection algorithm (NSA), a popular, prominent algorithm of the AIS domain based on the human immune system. IDS offer intrusion detection capabilities, both locally and in a network environment. The paper offers a review of recent solutions employing AIS in IDS, capable of detecting anomalous network traffic/breaches and operating system file infections caused by malware. A discussion regarding the reviewed research is presented with an analysis and suggestions for further research, and then the work is concluded.

Keywords
artificial immune systems, cybersecurity, intrusion detection, negative selection, malware

1. Introduction

IN the contemporary digital era, computer systems enjoy immense popularity. However, this widespread use has not come without drawbacks, as it has attracted actors with various motivations, many of whom frequently seek

unauthorised access to user data. The threat to computer systems doesn't just stem from individuals desiring to remotely control compromised workstations but is also posed by malicious software, commonly known as malware.

In response to these escalating threats, there has been significant development in intrusion detection systems (IDS) over the past few decades. These systems are dedicated to identifying and combatting both network and local infections, representing a crucial and rapidly evolving category of software within the cybersecurity domain. IDS employ a range of strategies for threat mitigation, which may include, for instance, the filtering of network packets based on predefined rules, or utilising a database of antivirus software signatures.

However, these traditional methods often fall short in detecting novel or previously unidentified threats. This limitation initiated the development of the first IDS inspired by artificial immune systems (AIS), conceptualised to overcome the constraints of their predecessors. IDS that incorporate AIS typically rely on algorithms, such as the negative selection algorithm (NSA) [1], positive selection [2], or clonal selection [3], all of which draw inspiration from biological immune systems. The need to study AIS in intrusion detection for cybersecurity arises from their adaptability and learning capabilities, which are crucial for countering evolving cyber threats. Unlike traditional systems that rely on known threat patterns, AIS can identify and adapt to new/unknown threats in a similar way to biological immune responses. These capabilities are especially vital in tackling zero-day attacks and advanced cyber threats that evade conventional detection methods. Also, the self-organising nature of AIS enables autonomous operation which may be essential in large-scale networking environments where manual monitoring is impractical. The ability of AIS to reduce false positives and their resilience against advanced evasion methods further highlights their suitability for modern applied cybersecurity.

Of the aforementioned algorithms, the NSA approach in particular has garnered substantial attention from the global scientific community. This algorithm functions by generating a collection of receptors, serving as the cyber equivalent of antibodies and T lymphocytes in a biological immune framework. The concept hinges on the principle that these digital "receptors" can identify and flag non-self elements, akin to how a living organism's immune system detects and responds to pathogens. This innovation marks a significant stride forward in ensuring cybersecurity by mimicking the resilience and adaptability of biological immune responses.

The generation of these receptors within the system can be accomplished through various methodologies: some may be randomly created [4], others might be pattern-based [5], among other techniques. Furthermore, these receptors operate based on a parameter known as the activation threshold. Depending on the specific implementation, this threshold may be fixed [6] or varying [7]. However, the application of these solutions has frequently encountered limitations, such as constraints related to the size of the processed files or the presence of vulnerabilities that result in a high percentage of undetectable infections.

To overcome these limitations, the past decade or so has witnessed the emergence of numerous modifications to the NSA, which incorporate various enhanced learning methods for training the receptor set. These methods include the use of real-values [8], Voronoi diagrams [9], two-stage training [10], hierarchical clustering [11], genetic algorithms [12], and mechanisms of adaptive immunoregulation [13]. These solutions are geared towards finding the most effective ways to train receptors, with a prevailing emphasis on approaches that employ variable activation thresholds. The result is an increasingly sophisticated system capable of processing large files and mitigating vulnerabilities.

The aim of this paper is to provide an overview of artificial immune systems used in intrusion detection systems, particularly the negative selection algorithm, and to provide a review of efforts in this field with regard to local and network applied cybersecurity.

2. Background

The fundamental concept that necessitates definition is the security of a computer system. But first, we need to describe what we mean by a computer system. A computer system is defined as an integrated set of hardware and software components that work together to enable users to perform specific computational tasks [14]. An individual instance of a computer system, which might be a personal computer or a high-performance setup for more demanding tasks, is often referred to as a computing device or simply a computer.

The components of a computer system are divided into physical and logical categories [14]. The physical category encompasses computer hardware like the motherboard, RAM, processor, and hard drive [15]. In contrast, the logical category involves various types of data and

software: this includes the configuration of physical components (such as UEFI/BIOS settings), system firmware, the operating system that manages the computer's functions, and user data stored on the hard drive [14, 15].

Computer system security denotes the system's resilience to various threats and unauthorised access [16]. The process of securing a system is a comprehensive effort directed towards the safeguarding of both the hardware components and the data contained within the computing device [16]. Thus, when we discuss computer system security, this encompasses the safety protocols for both physical hardware and the data processed and stored by the system.

Physical security is typically ensured by protecting the computing device from unauthorised physical access by third parties. Data security, however, is a multifaceted challenge. It is not solely dictated by the configuration of UEFI/BIOS, firmware, and the operating system. Instead, it includes a broader suite of protective measures. These can span data encryption, establishment of stringent access controls, network security measures, secure communication protocols, regular software updates, and the implementation of secure data storage and transmission practices.

Several fundamental aspects comprise computer system security [5]:

- **Availability** – a computer system should ideally provide uninterrupted access to its resources and data for authorised users,
- **Data confidentiality** – the system should ensure the confidentiality of user data, preventing access by unauthorised individuals,
- **Integrity** – data within the system should be protected against unwanted alterations, whether it is deletion, overwriting, or corruption,
- **Accurate threat classification** – security software should strive to minimise instances of false positive detections,
- **Accountability** – the computer system should have a built-in logging mechanism so that in the event of a security breach, it is possible to detect the incident and identify potential culprits.

Depending on certain factors, it might be essential to focus on specific aspects of security mentioned above. For instance, if a workstation

is used for data archiving, it might be crucial to concentrate on the integrity aspect of the system's data. A computer system's security policy is determined by the security aspects the system administrator focuses on [17].

An intrusion, or breach, refers to the act of unauthorised individuals accessing a computer system's data. It is important to note that "access" doesn't just refer to data viewing but also includes modifying or deleting them. In such instances, there is a violation of availability, data confidentiality, and system integrity principles. Intrusions can be committed directly by human actors (like hackers) or automated threats (using malware-type software). Breaches carried out by malicious software often serve as preliminary intrusions, paving the way for human actors to access data unauthorisedly [18].

The implementation of a chosen computer system security policy relies on selecting appropriate methods to address specific issues. To protect a workstation from intrusions, an administrator might employ software specialised to prevent such activities. For instance, securing the system against network intrusion attempts can begin with the installation of firewall software, allowing the administrator to block selected system network ports, among other things. This kind of blockade significantly hinders attacks on the ports specified by the administrator. A critical step in securing the system is installing software that detects malicious programs and network traffic. An IDS can constitute such software.

2.1. Intrusion detection systems

In recent years, tools known as intrusion detection systems (IDS) have gained significant traction within the scientific community. These tools are designed to differentiate between desirable and undesirable events through specific operational methods. In general, IDS are primarily employed for identifying unwanted network activities, but they can also serve to detect local threats [19].

IDS essentially utilise two basic primary techniques: rule-based detection and profile-based detection [20]. The former involves matching a sequence of samples against known patterns, which are identified as harmful, termed as "signatures". The latter, on the other hand, relies on system behaviour analysis to detect activities that deviate from the "normal" operational patterns of the environment. However, these fundamental IDS techniques do not incorporate dynamic learning or adaptation based on mitigated intrusions, limiting their capacity for advanced detection of unknown threats.

Consequently, researchers have been motivated to explore contemporary solutions that align with IDS themes. One such solution, derived from nature itself, is the biological immune system (BIS). This system consists of biological structures and processes within an organism that protect against diseases. For effective operation, the immune system must possess the capability to detect a wide array of harmful agents, such as viruses, bacteria and parasites, and distinguish them from the organism's healthy tissues [5].

The immune system is fundamentally composed of two subsystems: the innate system and the adaptive system. The innate system, present in nearly all living organisms, provides what is known as innate (or nonspecific) immunity. Its response to an invading pathogen is immediate; however, it does not retain memory of the exposure and, therefore, does not construct immunological memory [5]. In contrast, the adaptive system allows an organism to build immunological memory, providing specific immunity. Cells involved in this process include T lymphocytes and B lymphocytes, among others. T lymphocytes are responsible for cellular response, eliminating infected or mutated cells, while B lymphocytes are tasked with producing proteins called antibodies. These antibodies bind to antigens on the surfaces of harmful cells, effectively "marking" them for destruction, thereby initiating the humoral response [5]. This intricate biological framework provides an inspiration for cybersecurity measures, propelling the exploration and implementation of advanced and adaptive IDS strategies. One such strategy is artificial immune systems (AIS).

2.2. Artificial immune systems

The innovative field of artificial immune systems has emerged from the parallels between the functionalities of intrusion detection systems and biological immune systems. AIS encompass a suite of computational methods that draw inspiration from biological immunity, appealing due to their inherent capabilities for learning and adaptation within a given environment [21]. A pivotal feature of IDS strategies based on AIS algorithms is their proficiency in distinguishing "self" from "non-self" cells. Notably, algorithms widely applied in AIS-related matters include the negative selection algorithm [19], positive selection algorithm [22], and clonal selection algorithm [3]. These AIS algorithms share certain similarities with neural networks, as they incorporate system training based on a specified dataset.

The negative selection algorithm (NSA), inspired by the adaptive mechanisms of biological immune systems, operates by generating

binary strings that can match foreign strings while never aligning with self strings [23]. If a generated binary string matches a self string, it is discarded, mirroring the adaptive system's production of antibodies that bind only with harmful antigens and T lymphocytes that recognise only foreign cells [24].

Conversely, the positive selection algorithm (PSA) functions similarly to the NSA, but instead of matching foreign strings, the strings it produces align solely with self strings [2].

The clonal selection algorithm (CSA) is inspired by the biological immune response that triggers the proliferation of antibodies identifying a specific antigen. The activation of B lymphocytes (for particular antibodies) prompts their cloning, followed by intensive genetic mutation of the antibodies to enhance their antigen compatibility [25]. Similarly, the algorithm identifies the best-matching binary strings and clones them for further mutation, improving the compatibility of the mutated strings [25]. It is employed as a supplementary algorithm to the NSA and PSA.

Both the NSA and PSA present unique advantages and disadvantages. Research indicated in [22] suggests that detection efficiency is superior when using the positive selection algorithm. However, if the number of strings generated by the NSA is fewer than the number of self strings, the negative selection algorithm may prove more effective [7].

2.3. Negative selection algorithm

The primary objective of the negative selection algorithm (NSA) is to establish a collection of strings proficient in intrusion detection [6]. These strings, generated by the algorithm, are usually referred to as "receptors" or "detectors", but the term "antibodies" also occurs. Each receptor possesses a definitive length denoted as *l* [4]. *E*very prospective receptor undergoes scrutiny for its compatibility with any "self" string, wherein a "self" string signifies a sequence that should never be flagged as an anomaly. For this purpose, the NSA employs a parameter, *m,* representing the receptor's activation threshold [26]. A receptor is deemed activated if it matches another binary string. Depending on the rule used for matching, the matching process usually involves the occurrence of *m* identical, consecutive bits at the same position *k* in both the receptor and the binary string under examination [5].

If a generated string matches with at least one "self" string, it cannot become a receptor and is consequently dismissed. Traditionally, the

NSA assumes the existence of a single receptor set **R**, encompassing all generated receptors [4].

The methodology behind receptor generation is not predefined – many methods of generation exist – but for the sake of simplicity, the random generation method will be outlined [4]. Random generation involves a parameter R_{max}, which dictates the maximum count of receptors to be generated. Given a defined parameter *l*, R_{max} candidates for receptors are generated. Each receptor candidate is subjected to the aforementioned verification before being included in the resultant set **R**. Typically, receptors generated via NSA do not facilitate a 100% anomaly detection rate. Zones not covered by receptors are referred to as holes [27].

The negative selection algorithm can be adapted as the foundational mechanism for infection detection in IDS. When the NSA is employed for infection detection, the input from the receptor generator is substituted with a stream of strings for IDS examination. The set of "self" strings is replaced by the receptor set **R**. Compatibility is assessed using the same parameters *l* and *m* as in the case of receptor generation. If at least one "self" string matches, the algorithm ceases operation, signalling an infection detection, which is a divergence from the receptor-generation stage (where the algorithm would have rejected the receptor candidate instead).

Key performance indicators for the IDS and the algorithms applied within it, including the negative selection algorithm, are:

- **TP** (True Positives) – the count of accurately identified infections,
- **TN** (True Negatives) – the count of correctly unidentified infections,
- **FP** (False Positives) – the count of inaccurately identified infections,
- **FN** (False Negatives) – the count of inaccurately unidentified infections.

Additional indicators may be:

- the duration required for receptor generation,
- the quantity of receptors retained in memory following generation,
- memory usage by primary receptors,
- memory usage by all receptors,
- memory occupied by the original program.

3. Review of the use of artificial immune systems

Algorithms of artificial immune systems are eagerly employed, explored, and refined within the scientific community.

González, Dasgupta and Kozma [28] applied a data representation using a two-dimensional plane and real numbers for the space of self and non-self strings in their examination of the algorithm (RNSA – real-valued negative selection algorithm). Research was also conducted using binary numbers represented in Grey code. A crucial conclusion drawn by the researchers was the emphasis on the importance of appropriately tailoring the matching rule of the negative selection algorithm in accordance with the intended use of the IDS. They highlighted that for applications where the entire space of self strings is known (such as, for instance, scanning for data integrity verification), the generalisation of self-data is not as critical. Ji and Dasgupta [29] discussed the challenges encountered when implementing the NSA grounded in real number values. They posited that the majority of the problems reported are often the result of incorrect application of the technique or challenges that aren't exclusively related to negative selection algorithms. They argued that, in contrast, tests using artificial and established real-world data demonstrate that NSAs possess significant adaptability in maintaining a balance between effectiveness and robustness, as well as in incorporating elements tailored to specific fields within the approach, such as different types of distance calculations.

Ji and Dasgupta [7] enhanced the NSA through the introduction of variable-length detectors (V-detectors). These detectors, thanks to their variable length, more efficiently "plug" the holes that arise during generation. Studies demonstrated that the algorithm's performance improved without a significant increase in its complexity. Lu, Zhang, Wang, and Gong [30] proposed an NSA method using V-detectors for ransomware detection. In work [11], a fast negative selection algorithm based on the hierarchical structure of the self-string set was presented. Zhu, Chen, Yang, Li, Yang, and Zhang [31] utilised Voronoi diagrams to enhance the NSA. Their proposed VorNSA algorithm constructs a Voronoi diagram based on a test set, subsequently generating two types of receptors based on this diagram, reducing the receptor-generation time. The testing (detection) phase was also redesigned – data are divided into smaller intervals, mapped, and sorted during the reduction stage. Another approach using Voronoi diagrams is described in [9].

González, Dasgupta, and Niño [32] introduced a version of the negative selection algorithm, which was expanded to estimate the

optimal number of receptors needed to cover the space of non-self strings (RRNSA – randomised real-value negative selection algorithm). In addition to expanding the algorithm itself, the authors conducted an in-depth theoretical analysis forming the basis for performance analysis of their improved version. They inferred that the RRNS variant operates faster than RNS but noted that in certain cases, heuristic algorithms are even more efficient, although other algorithms may have a better theoretical foundation.

Marciniak, Wawryn and Widuliński [33] demonstrated the use of the negative selection algorithm for controlling a heating boiler. In [10], a version of the algorithm trained multiple times for a different number of self strings to enhance performance was described. The approach proposed in [13] takes into account the use of an adaptive immune regulation mechanism to calculate the radius on the plane of self strings. Saurabh and Verma [34] proposed an NSA version with a tuning function called NIIAD. Balicki [35] introduced NSA to overcome the limitations of a multi-criteria evolutionary algorithm. Study [36] indicated that AIS could be applied to threat detection in mobile operating systems.

In [37], a system called MILA (multilevel immune learning algorithm) was proposed, which considers not only the application of NSA but also receptor expansion using the clonal method and a dynamic receptor-generation method in one solution. Fakhari and Moghadam [38] introduced an NSA version named NSSAC, which is capable of adapting to data sets. Gao, Ovaska, and Wang [12] proposed a receptor-generation method based on a genetic algorithm in their work. Paper [39] describes an IDS system based on an evolutionary algorithm for anomaly detection in distributed computer systems. In [40], information about estimating the range of receptors in NSA was provided.

Kamal and Bhusry [41] presented negative selection algorithms optimised by artificial bee colonies (ABC algorithm). Nunes de Castro and von Zuben [42] described the aiNet system based on AIS algorithms for data analysis. Prathyusha and Kannayaram [43] introduced a novel mechanism based on AIS for mitigating DDoS network attacks in the cloud.

3.1. Use of artificial immune systems in intrusion detection

The use of artificial immune systems in intrusion detection systems is a popular notion among researchers.

In [44], the authors introduced an implementation of a clonal-based artificial immune system as the central mechanism for a network intrusion detection system.

The research was structured around two main stages: training and testing. The initial step in the training phase involved creating a series of "antibodies". These antibodies are essentially pieces of information that were derived from six specific types of network attacks: Smurf, Land, Satan, Neptune, Ipsweep, and Portsweep. Each antibody possesses eight unique features that allow it to effectively differentiate between these various forms of attacks: the duration, type of protocol, type of service, flag, source bytes, number of access files, number of outbound commands and service difference host rate.

In the testing phase, the researchers examined the effectiveness of their AIS-based IDS against the six types of attacks mentioned. The aim was to evaluate how well the system could detect these intrusions in practice, reflecting real-world applications where an IDS needs to reliably identify any attempt to breach network security.

The researchers used a dataset known as the KDD Cup, containing 284,948 connection data, of which 10% (28,494 connections) were randomly chosen for testing, while the rest were used for training. Initially, a probability value of 0.2 was employed, indicating a 20% chance of each attack connection being chosen for testing. The AIS algorithm correctly identified 27,552 out of 28,494 attack connections, a true-positive rate of roughly 97%.

Further experiments were conducted with different probability values (0.3, 0.4, and 0.5) to discern their effect on the study. The findings revealed that the AIS algorithm recognised more attack connections as the probability value increased. A 0.3 probability yielded a 97% true-positive rate, 0.4 resulted in 98%, and 0.5 demonstrated the highest rate at 99.86%, with only 39 attack connections not correctly identified.

The authors observed that using a high probability value for selection might skew the testing dataset towards connections from the early part of the dataset, possibly consisting of many similar data connections, since the same attack data are grouped together in the raw dataset. This could reduce the effectiveness of testing the algorithm's performance in network intrusion detection. Hence, a smaller probability value is recommended to ensure a more even distribution of attack patterns in the testing dataset. Regarding the

training process, the primary aim was to generate antibodies with high fitness values that are considered crucial for recognising attack data during testing. The fitness value in this context ranges between -1 and 1, with values close to 1 indicating high-quality antibodies. The AIS algorithm, after running 100 iterations, produced the best-quality antibody with a fitness value of 0.46 using a 0.2 selection probability. Other probabilities yielded slightly lower fitness values, with the 0.5 probability producing the lowest-quality antibody with a fitness value of just 0.41.

The authors concluded that the cloning and mutation processes are crucial for the suggested algorithm to produce effective solutions during training. The positive results shown by AIS demonstrated its capability to address the issue, matching the performance of other methods in existing scientific research.

Study [45] introduces an Internet of Things (IoT) anomaly intrusion detection system specifically for smart homes, employing a hybrid model that combines artificial immune system and extreme learning machine (ELM) methodologies, referred to as the AIS-ELM IDS framework. This system is integrated into a smart home environment through a Mozilla gateway installed on a Raspberry Pi, which connects all smart devices via a router using the REST API for streamlined monitoring and control.

The AIS component of the IDS uses the clonal selection method to enhance the system's detection capabilities through receptor maturation. The process begins with an initialisation stage where input data is assessed to determine the optimal inputs with the highest affinity and lowest negative selection. This is followed by the clonal selection stage, encompassing clonal, mutation, and substitution phases.

The ELM algorithm assigns arbitrary input weights and biases, calculates a hidden layer output matrix, and determines the output weight. The integration of AIS and ELM processes in the IDS helps in the accurate detection of normal and abnormal patterns in network traffic, flagging them as "1" for normal and "0" for anomalies.

The system enhances home security by initiating an immediate response when an anomaly is detected. It employs a custom-designed alarm system to alert the homeowner, prompting them to act – either by disconnecting the internet in the event of an external threat or by isolating the compromised segment within the smart home for internal threats. If the system doesn't detect any user action within two minutes, it autonomously disconnects the internet, adding an

extra layer of security. This approach not only optimises intrusion detection but also provides an automated, rapid response mechanism.

Brown, Anwar and Dozier [46] proposed the modified artificial immune system (mAIS) model. In mAIS, two usual sets of detectors are developed: the self detector set and the non-self detector set. Following generation of the detector sets, in the "Proportion Based Classification" phase, these detector sets work in tandem to classify unknown traffic instances. An instance is labelled as non-self or abnormal if a larger fraction of non-self detectors identify it compared to self detectors, and vice versa. Given the potentially more severe consequences of false negatives compared to false positives, any instance equally identified by both detector types is classified as non-self to minimise risk.

An "Interval Matching Rule" is employed, involving each detector's 41 intervals, each corresponding to a specific dataset feature. A match between a detector and an instance is determined by selecting an "r-value". If the number of features within a detector's intervals meets or exceeds this r-value, the detector is considered to match the instance.

The dataset used for this work was the UNB ISCX Intrusion Detection Evaluation Dataset, selected for its recentness and relevance to contemporary network scenarios. It comprises 148,517 instances of network traffic, with 77,054 normal and 71,463 anomalous instances. Each instance has 41 different features. The testing method involved dividing the dataset into training, tuning, and test sets. Initially, all instances were in the training set, from which 50,000 instances were moved into the test set and another 50,000 into the tuning set, leaving 48,517 in the training set. This process was repeated 30 times for each of the 6 folds. Detectors were evaluated and those not matching any instance were promoted to mature detectors. The best-performing detectors were retained for testing. After each fold, data sets were rotated and the process repeated, resulting in 180 total runs.

The study opted for 1,000 initial immature detectors to reduce computational demands, with a fixed width of 1.0 for the detectors. While general detectors cover more hypothesis space, they can increase false positives.

Experiments ran for approximately 10 hours on a test computer. As per the results, the standard AIS marginally surpassed the mAIS in detection rate and accuracy, whereas the mAIS performed slightly

better in terms of the true negative rate and false positive rate. The standard AIS covered more hypothesis space due to less internal competition between detectors. The authors concluded that both standard AIS and MAIS demonstrated similar performance levels on the dataset utilised. The varied nature of normal and abnormal network instances might contribute to this outcome, potentially restricting the efficiency of MAIS. The authors suggested that employing a larger and more diverse set of initial detectors could enhance the performance of both systems.

Tosin and Gbenga [47] enhanced their proposed network intrusion detection system by integrating the NSA with a feature selection mechanism. Due to the NSA's non-prior knowledge requirement and nature as a one-class classifier, NSA faces scalability issues due to the large number of detectors needed and high false positives. To address the scalability issues, the research introduces a feature selection process, utilising an artificial neural network (ANN) to reduce the dimensionality of the input data, thereby tackling NSALG's scalability issue. This process involves passing each feature (data column) through the ANN to evaluate its relevance based on classification accuracy, with those exceeding 80% accuracy being retained. The methodology encompasses three stages: data preprocessing for normalisation and feature selection, the NSA stage for detector generation and anomaly detection, and finally, an alert generation phase.

Utilising the NSL-KDD dataset, the model's performance was evaluated using a confusion matrix approach. The experiments were conducted in two scenarios: with and without the feature selection mechanism. Improvements were observed when the feature selection was employed. Specifically, there were significant increases in true-positive rate (TPR), true negative rate (TNR), and overall accuracy (ACC), alongside reductions in the false positive rate (FPR) and false negative rate (FNR). TPR saw an 11.65% increase, TNR improved by 213.91%, and ACC increased by 26.54%. FPR and FNR decreased by 70.62% and 19.75%, respectively, indicating fewer false alarms and missed detections.

In the work's conclusion, the authors state that the integration of feature selection with NSA substantially enhanced IDS performance by mitigating scalability issues.

Local-based intrusion detection systems utilising AIS also exist. In strona 73, Widuliński and Wawryn explored the possibility of employing an AIS-based IDS locally to scan for infections on a computer. They discuss an advanced system for detecting unauthorised

changes to files within an operating system. The IDS works by constantly monitoring a designated area within the operating system, which the user sets up first. Its primary job is to scan files in this area to detect any unexpected or suspicious alterations that are indicative of potential security threats or malware intrusions. The IDS's functionality is managed by a central component called the control unit (CU). The CU oversees the operations of two critical parts of the system: the receptor-generation unit (RGU) and the anomaly detection unit (ADU). When the system starts, the RGU runs first. Its role is to create the set(s) of receptors which will be used to identify whether the system's files have been tampered with.

In the case of this IDS, these receptors are binary strings, sequences of bits: ones and zeros, with a specific length. They're designed to detect "non-self" data – essentially, infections or modifications – within the monitored program files. Each file under surveillance gets its own unique set of receptors, which are stored separately, either in memory (RAM) or as a file on non-volatile storage such as a hard drive or flash drive, to ensure they're secure and intact. The system is designed with a special interface to allow the IDS to be adaptable and functional across different platforms.

Once the receptors are generated, the CU instructs the ADU block to start operation. The ADU scans the safeguarded files, comparing their contents with the receptors. This comparison is done using a formula (or rule) that checks for matching bit patterns between the receptors and each 32-bit segment of the monitored program's bytes. When a match is found, it flags that part of the program as potentially compromised.

In instances where the ADU identifies an intrusion, it logs the issue. Afterwards, it informs the user precisely where the problem is and what parts of the data have been altered – likely due to malicious software (malware). The system doesn't stop after one scan; it continues to check the files repeatedly until the user decides to halt operation. However, legitimate updates to files, such as when a software update occurs, necessitate the creation of new receptors. If there is a valid change, the system doesn't mistake it for an intrusion; the CU simply instructs the RGU to start the receptor-generation process anew. In strona 73, a modification of the NSA was proposed to mitigate false negatives when anomalies (or infections) occurred between 32-bit program memory cells. The modification, called intercellular receptors (ICR), offers an additional, smaller receptor set to assist with detection of infections that might occur between memory cells.

4. Discussion

The reviewed research highlights the versatility of artificial immune systems, particularly when used with intrusion detection systems, which is a domain of cybersecurity. The adaptive and self-learning characteristics of AIS algorithms have shown considerable promise in identifying and responding to network intrusions, underlining their adaptability and efficiency in real-world applications.

Research by González, Dasgupta, and others highlights the importance of tailoring the matching rule in negative selection algorithms for specific applications, affirming that the flexibility of AIS can be optimal when the algorithms are adapted for their intended purposes. The introduction of variable-length detectors, as discussed by Ji and Dasgupta, and the use of V-detectors in the algorithms, show an evolutionary leap, enhancing detection efficiency without substantially increasing system complexity.

The effectiveness of AIS in IDS, as evidenced in studies [44] and [45], is particularly noteworthy. The high true-positive rates reported confirm the system's robustness and ability to identify network intrusions. However, the studies also caution about potential biases in the testing dataset and the importance of a balanced and diverse set of data for training, highlighting that the reliability of AIS is significantly influenced by the quality of the input it receives. This is a critical insight, reflecting the principle that the output is only as good as the input.

Moreover, the integration of AIS with other methodologies, such as the extreme learning machine (ELM) in [45] and artificial neural networks (ANN) in [47], points towards a growing trend of hybrid models. These models aim to combine the strengths of various systems to achieve higher efficiency and reliability, while also addressing inherent challenges such as scalability issues and high false positives in NSA.

Despite these advances, studies such as those carried out by Brown, Anwar, and Dozier [46] suggest that there is still room for improvement, especially concerning the reduction of false positives and enhancement of detection accuracy. This indicates that while AIS solutions are a powerful tool, their efficacy can be further optimised, potentially through the integration of more diverse detectors, refinement of algorithms, or hybridisation with other effective techniques.

The IDS proposed by Widuliński and Wawryn introduces a localised solution for detecting unauthorised alterations within an operating

system. This approach represents an application of AIS in cybersecurity, marking a departure from more generic, network-focused IDS. The system's capacity to continually generate and update receptors allows for an adaptability and sensitivity to changes within the system's files. The operation of the IDS seems to face some unique challenges, particularly concerning the differentiation between legitimate alterations, like software updates, and unauthorised changes. The system's reliance on user-initiated receptor regeneration following legitimate updates could potentially introduce vulnerabilities, particularly if the user is unaware of the necessity of this action following updates. The introduction of intercellular receptors (ICR) addresses a critical gap in traditional locally utilised NSA methodologies by targeting the detection of anomalies occurring between 32-bit memory cells, and improves the true-positive rates by about 15% while slightly increasing the memory usage.

Reviewing the recent advances in local and network AIS-based cybersecurity, a distinct lack of IDS solutions combining both local and network anomaly detection can be observed. A novel hybrid AIS-based IDS that integrates both local and network detection capabilities would represent a significant advancement in cybersecurity. Such a system could combine the strengths of both approaches to provide a more comprehensive defence mechanism against a variety of cyber threats. Some of the potential benefits of such a system could include:

- **Dynamic receptor generation** – the system could continuously update its defence mechanisms based on new potential threats detected across the network and local machines. This would be especially beneficial in combatting zero-day exploits, where traditional signature-based methods are inadequate,
- **Context-aware detection** – by analysing data from both the local environment and network traffic, the hybrid IDS could employ machine learning algorithms to better understand the context, enhancing its ability to distinguish between normal changes and potential threats,
- **Real-time cross-verification** – when an anomaly is detected locally, the system could cross-verify it with network data to confirm if the anomaly is an isolated incident or part of a broader network intrusion,
- **Adaptive learning** – over time, the hybrid system could learn from the traffic patterns and typical file changes within the network and local systems, improving its detection rates further.

Nonetheless, the development of such a hybrid system would also pose some challenges, such as the complexity of integrating local and network IDS functionalities, potential privacy concerns, and the increased system resources required.

The overview presented in this paper contributes to the recent state of research in the field of cybersecurity by offering a focused analysis of the NSA within AIS for IDS. A detailed exploration of the NSA's theoretical and practical applications, highlighting recent advancements, has been provided. The interdisciplinary approach – drawing insights from biological systems – also highlights the connection between biology and cybersecurity, encouraging innovative ideas in IDS research.

While a comprehensive overview of the application of AIS in IDS has been provided, it is also important to acknowledge certain limitations inherent in this focused approach. The primary limitation is the concentrated emphasis on NSA. While NSA is a significant and influential algorithm within AIS, the focus on this single algorithm potentially overlooks the diverse range of other algorithms within the AIS domain, such as the positive selection algorithm (PSA) or the clonal selection algorithm (CSA). This narrow scope may limit the comprehensiveness of the review in capturing the full spectrum of AIS capabilities. Another limitation of the work is the lack of a comparative analysis with non-AIS-based IDS approaches, which would be adequate for providing a balanced view of where NSA stands in relation to other methodologies.

5. Conclusions

An overview of artificial immune systems, intrusion detection systems and a review of efforts in the field have been presented. The reviewed research shows significant potential of AIS in enhancing intrusion detection systems. The adaptability, versatility, and self-regulatory aspects of AIS make it a formidable approach to securing local computers and networks against a variety of intrusions.

In conclusion:

- Tailoring algorithms to specific applications enhances the effectiveness of AIS. This customisation, particularly in negative selection algorithms, is crucial for optimising performance in different environments.

- The introduction of innovative methods, such as variable-length detectors and the use of Voronoi diagrams, improves the

efficiency of intrusion detection without overly complicating the systems.

- Hybrid models, combining AIS with other techniques like ELM or ANN, have emerged as highly effective in improving accuracy and reducing false positives, indicating a promising direction for future research and application.

- Despite the demonstrated efficacy of AIS in IDS, there remains a need for further refinement to reduce false positives and improve detection accuracy.

- The success of AIS significantly hinges on the quality of data used for training, stressing the importance of proper datasets that reflect real-world scenarios.

All in all, AIS hold substantial promise in the realm of IDS, providing a robust, adaptable, and intelligent approach to local and network cybersecurity. Continued research and development in this field are to be encouraged, focusing on customised solutions, algorithmic advancements, and hybrid models, to fully realise the potential of AIS in safeguarding digital environments. Research on hybrid solutions combining local and network approaches in particular appears to be a reasonable avenue to explore in the future.

References

[1] P. Helman and S. Forrest, "An efficient algorithm for generating random antibody strings," Technical Report cs-94-07, The University of New Mexico, 1994.

[2] H. Alrubayyi, G. Goteng, M. Jaber, and J. Kelly, "A Novel Negative and Positive Selection Algorithm to Detect Unknown Malware in the IoT," in IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 1-6, 2021. doi: 10.1109/infocomwkshps51825.2021.9484483.

[3] A. S. Perelson and G. F. Oster, "Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination," *Journal of Theoretical Biology*, vol. 81, no. 4, pp. 645–670, 1979, doi:10.1016/0022-5193(79)90275-3.

[4] S. Forrest, A. S. Perelson, L. Allen, and R. Cherukuri, "Self-nonself discrimination in a computer," *Proceedings of 1994 IEEE Computer Society Symposium on Research in Security and Privacy*, pp. 202–212, 1994, doi:10.1109/risp.1994.296580.

[5] S. Hofmeyr, "An Immunological Model of Distributed Detection and Its Application to Computer Security," doctoral dissertation, University of Witwatersrand, Johannesburg, South Africa, 1999.

[6] D. Li, S. Liu, and H. Zhang, "Negative selection algorithm with constant detectors for anomaly detection, "*Applied Soft Computing*, vol. 36, pp. 618–632, 2015, doi:10.1016/j.asoc.2015.08.011.

[7] Z. Ji and D. Dasgupta, "Estimating the detector coverage in a negative selection algorithm," *Proceedings of the 2005 conference on Genetic and evolutionary computation – GECCO '05*, pp. 281–288, 2005, doi:10.1145/1068009.1068056.

[8] S. E. Dixon, "Studies on Real-Valued Negative Selection Algorithms for Self-Nonself Discrimination," M. Sc. thesis, California Polytechnic State University, San Luis Obispo, USA, 2010.

[9] G. Zhao et al., "Voronoi-Based Continuous k Nearest Neighbor Search in Mobile Navigation," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 6, pp. 2247–2257, 2011, doi:10.1109/tie.2009.2026372.

[10] M. Gong, J. Zhang, J. Ma, and L. Jiao, "An efficient negative selection algorithm with further training for anomaly detection," *Knowledge-Based Systems*, vol. 30, pp. 185–191, 2012, doi:10.1016/j.knosys.2012.01.004.

[11] W. Chen, T. Li, X. Liu, and B. Zhang, "A negative selection algorithm based on hierarchical clustering of self set," *Science China Information Sciences*, vol. 56, no. 8, pp. 1–13, 2011, doi:10.1007/s11432-011-4323-7.

[12] X. Gao, S. Ovaska, and X. Wang, "Genetic Algorithms-based Detector Generation in Negative Selection Algorithm," *2006 IEEE Mountain Workshop on Adaptive and Learning Systems*, pp. 133–137, 2006, doi:10.1109/smcals.2006.250704.

[13] H. Deng and T. Yang, "A negative selection algorithm based on adaptive immunoregulation," *2020 5th International Conference on Computational Intelligence and Applications (ICCIA)*, pp. 177–182, 2020, doi:10.1109/iccia49625.2020.00041.

[14] A. Elahi, C*omputer Systems: Digital Design, Fundamentals of Computer Architecture and Assembly Language*, 1st ed. Cham: Springer, 2018.

[15] N. Nisan and S. Schocken, *The Elements of Computing Systems: Building a Modern Computer from First Principles*, 1st ed. Cambridge, MA: The MIT Press, 2005.

[16] L. F. Reese, "Challenges faced today by computer security practitioners," *[1989 Proceedings] Fifth Annual Computer Security Applications Conference*, 1989.,doi:10.1109/csac.1989.81044.

[17] L. Mixia, Y. Dongmei, Z. Qiuyu, and Z. Honglei, "Network Security Risk Assessment and Situation Analysis," *2007 International Workshop on Anti-Counterfeiting, Security and Identification (ASID)*, 2007, doi:10.1109/iwasid.2007.373676.

[18] A. Datta, S. Jha, N. Li, D. Melski, and T. Reps, "Analysis Techniques for Information Security," *Synthesis Lectures on Information Security, Privacy, and Trust*, vol. 2, no. 1, pp. 1–164, 2010, doi:10.2200/s00260ed1v01y201003spt002.

[19] C. J. Delona, P. V. Haripriya, and J. S. Anju, "Negative Selection Algorithm: A Survey," *International Journal of Science, Engineering and Technology Research (IJSETR)*, vol. 6, no. 4, pp. 711–715, 2017.

[20] L. Reznik, "Intrusion Detection Systems," in *Intelligent Security Systems: How Artificial Intelligence, Machine Learning and Data Science Work For and Against Computer Security*, 1st ed. Hoboken, NJ: Wiley-IEEE Press, 2022, pp. 109–176.

[21] J. D. Farmer, N. H. Packard, and A. S. Perelson, "The immune system, adaptation, and machine learning," *Physica D: Nonlinear Phenomena*, vol. 22, no. 1–3, pp. 187–204, 1986, doi:10.1016/0167-2789(86)90240-x.

[22] F. Zhang and Y. Ma, "Integrated Negative Selection Algorithm and Positive Selection Algorithm for malware detection," *2016 International Conference on Progress in Informatics and Computing (PIC)*, pp. 605–609, 2016, doi:10.1109/pic.2016.7949572.

[23] M. Ayara, J. Timmis, R. de Lemos, L. N. de Castro, and R. Duncan, "Negative selection: How to generate detectors," *Proceedings of the 1st International Conference on Artificial Immune Systems (ICARIS)*, pp. 182–196, 2002.

[24] R. J. De Boer and A. S. Perelson, "How diverse should the immune system be?," *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 252, no. 1335, pp. 171–175, 1993, doi:10.1098/rspb.1993.0062.

[25] L. N. de Castro and F. J. Von Zuben, "Learning and optimization using the clonal selection principle," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 3, pp. 239–251, 2002, doi:10.1109/tevc.2002.1011539.

[26] F. González, D. Dasgupta, and J. Gómez, "The Effect of Binary Matching Rules in Negative Selection," *Genetic and Evolutionary Computation – GECCO 2003*, pp. 195–206, 2003, doi:10.1007/3-540-45105-6_25.

[27] P. D'haeseleer, S. Forrest, and P. Helman, "An immunological approach to change detection: algorithms, analysis and implications," *Proceedings 1996 IEEE Symposium on Security and Privacy*, 1996, doi:10.1109/secpri.1996.502674.

[28] F. Gonzalez, D. Dasgupta, and R. Kozma, "Combining negative selection and classification techniques for anomaly detection," *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)*, pp. 705–710, 2002, doi:10.1109/cec.2002.1007012.

[29] Z. Ji, "Negative selection algorithms: From the thymus to V-detector," PhD dissertation, Department of Computer Science, The University of Memphis, Memphis, Tennessee, USA, 2006.

[30] T. Lu, L. Zhang, S. Wang, and Q. Gong, "Ransomware detection based on V-detector negative selection algorithm," *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pp. 531–536, 2017, doi:10.1109/spac.2017.8304335.

[31] F. Zhu, W. Chen, H. Yang, T. Li, T. Yang et al., "A Quick Negative Selection Algorithm for One-Class Classification in Big Data Era," *Mathematical Problems in Engineering*, vol. 2017, pp. 1–7, 2017, doi:10.1155/2017/3956415.

[32] F. González, D. Dasgupta, and L. F. Niño, "A Randomized Real-Valued Negative Selection Algorithm," *Lecture Notes in Computer Science*, vol. 2787, pp. 261–272, 2003, doi:10.1007/978-3-540-45192-1_25.

[33] J. Marciniak, K. Wawryn, and P. Widulinski, "An artificial immune negative selection algorithm to control water temperature in the outlet of the chamber," *2018 International Conference on Signals and Electronic Systems (ICSES)*, pp. 236–241, 2018, doi:10.1109/ICSES.2018.8507293.

[34] P. Saurabh and B. Verma, "A Novel Immunity inspired approach for Anomaly Detection," *International Journal of Computer Applications*, vol. 94, no. 15, pp. 14–19, 2014, doi:10.5120/16418-6034.

[35] J. Balicki, "Negative Selection with Ranking Procedure in Tabu-Based Multi-criterion Evolutionary Algorithm for Task Assignment," *Computational Science – ICCS 2006*, pp. 863–870, 2006, doi:10.1007/11758532_112.

[36] J. Brown, M. Anwar, and G. Dozier, "Detection of Mobile Malware: An Artificial Immunity Approach," *2016 IEEE Security and Privacy Workshops (SPW)*, pp. 74–80, 2016, doi:10.1109/spw.2016.32.

[37] D. Dasgupta, "Immunity-based Intrusion Detection System: A General Framework," *Proceedings of 22nd National Information Systems Security Conference*, pp. 147–160, 1999.

[38] S. N. S. Fakhari and A. M. E. Moghadam, "NSSAC: Negative selection-based self adaptive classifier," *2011 International Symposium on Innovations in Intelligent Systems and Applications*, pp. 29–33, 2011, doi:10.1109/inista.2011.5946064.

[39] C. R. Haag, G. B. Lamont, P. D. Williams, and G. L. Peterson, "An artificial immune system-inspired multiobjective evolutionary algorithm with application to the detection of distributed computer network intrusions," *Proceedings of the 2007 GECCO conference companion on Genetic and evolutionary computation – GECCO '07*, pp. 420–435, 2007, doi:10.1145/1274000.1274035.

[40] Z. Ji, D. Dasgupta, "Real-Valued Negative Selection Algorithm with Variable-Sized Detectors," *Genetic and Evolutionary Computation – GECCO 2004*, pp. 287–298, 2004, doi:10.1007/978-3-540-24854-5_30.

[41] P. Kamal and M. Bhusry, "Artificial Bee Colony Optimization based Negative Selection Algorithms to Classify Iris Plant Dataset," *International Journal of Computer Applications*, vol. 133, no. 10, pp. 40–43, 2016, doi:10.5120/ijca2016908072.

[42] L. Nunes de Castro and F. J. Von Zuben, "aiNet: An Artificial Immune Network for Data Analysis," *Data Mining: A Heuristic Approach*, pp. 231–260, 2002, doi:10.4018/978-1-930708-25-9.ch012.

[43] D. J. Prathyusha and G. Kannayaram, "A cognitive mechanism for mitigating DDoS attacks using the artificial immune system in a cloud environment," *Evolutionary Intelligence*, vol. 14, no. 2, pp. 607–618, 2020, doi:10.1007/s12065-019-00340-4.

[44] S. I. Suliman, M. S. Abd Shukor, M. Kassim, R. Mohamad, and S. Shahbudin, "Network Intrusion Detection System Using Artificial Immune System (AIS)," *2018 3rd International Conference on Computer and Communication Systems (ICCCS)*, pp. 178–182, 2018, doi:10.1109/CCOMS.2018.8463274.

[45] E. D. Alalade, "Intrusion Detection System in Smart Home Network Using Artificial Immune System and Extreme Learning Machine Hybrid Approach," *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, pp. 1–2, 2020, doi:10.1109/WF-IoT48130.2020.9221151.

[46] J. Brown, M. Anwar and G. Dozier, "Intrusion Detection Using a Multiple-Detector Set Artificial Immune System," *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, pp. 283–286, 2016, doi:10.1109/IRI.2016.45.

[47] S.-I. T. Tosin and J. R. Gbenga, "Negative selection algorithm based intrusion detection model," *2020 IEEE 20th Mediterranean Electrotechnical Conference (MELECON)*, pp. 202–206, 2020.

[48] P. Widulinski and K. Wawryn, "A human immunity inspired intrusion detection system to search for infections in an operating system," *2020 27th International Conference on Mixed Design of Integrated Circuits and Systems (MIXDES)*, pp. 187–191, 2020, doi:10.23919/MIXDES49814.2020.9155771.

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Shielding the Spanish Cyberspace: An Interview with Spain's National Cryptologic Centre (CCN)

Rubén Arcos | University Rey Juan Carlos, Madrid, Spain, ORCID: 0000-0002-9665-5874

Abstract

This interview between Rubén Arcos and Spain's National Cryptologic Centre (CCN) was conducted via email on 24 October 2022. CCN is part of Spain's National Intelligence Centre (CNI), and through its national alert and response centre against cyberattacks and cyber threats, CCN-CERT, it contributes to the cybersecurity of Spain. The discussion focuses on Spain's approach to cybersecurity, existing tools for information sharing/management of cyber incidents and tools supporting the production of intelligence on cyber threats. It also deals with current and emerging trends in the cyber domain and developments and activities in the fields of prevention, detection and response. Finally, the interview highlights measures in the March 2022 National Cybersecurity Plan and initiatives against potential cyber-attacks during elections.

Received: 16.12.2022

Accepted: 10.02.2023

Published: 13.02.2023

Cite this article as:
R. Arcos, „Shielding the Spanish Cyberspace: An Interview with Spain's National Cryptologic Centre (CCN)," ACIG, vol. 2, no. 1, 2023. DOI: 10.5604/01.3001.0016.2484

Corresponding author:
Rubén Arcos, University Rey Juan Carlos, Madrid, Spain; ORCID: 0000-0002-9665-5874; E-MAIL: ruben.arcos@urjc.es

Spain is very well-ranked in the last edition of the Global Cybersecurity Index of the International Communication Union (ranked 4th together with the Republic of Korea and Singapore) and is within the three top-ranked countries for Europe region[1]. How would you describe the Spanish model or approach to cybersecurity and the reasons that have led to achieving these results?

The Spanish approach is described in the attached document (https://www.ccn.cni.es/index.php/es/menu-ccn-es/aproximacion-espanola-a-la-ciberseguridad), but there are aspects highlighted below in which we believe our approach is leading in the European Union and elsewhere:

- A mandatory National Cybersecurity Framework for the public sector that obliges the application of 7 basic principles, 15 minimum requirements and, depending on the categorisation of the system, up to 73 cybersecurity measures. The framework is developed through more than 90 guidelines. It is also flexible in its application to smaller organisations ("PILAR" LEGAL MEASURES);
- An exchange of cyber incidents (more than 15,000 by 2022) across the public sector, which in turn allows for the implicit exchange of IOCs (Indicators of Compromise) and an automatic distribution of cyber intelligence tailored to the needs of the agencies ("PILAR" COOPERATIVE MEASURES);
- Common services for the development of guidelines and standards, as well as training courses;
- Articulation of a national network of SOCs (Security Operations Centers) as a result of previous experience;
- Use of common and interchangeable tools to enforce interoperability in cybersecurity ("PILAR" CAPABILITY DEVELOPMENT).

Which measures, if any, in your opinion would further improve the commitment of Spain to cybersecurity? And how do these excellent outcomes have an impact in deterring attacks from different actors in cyberspace against Spain and its national interests?

In this question we will focus on technical measures, but above all on the smooth exchange of cyber incidents and cyber threats. For this purpose, we use two platforms which, due to their flexibility, allow this exchange between organisations using CCN-CERT as a central exchange point.

1 Spain top-scores in 3 out of 5 pillars of the framework (legal measures, capacity development and cooperative measures) and has some room for improvement in the pillars of technical measures and organisational measures. See: https://www.itu.int/epublications/publication/DSTR-GCI.01-2021-HTM-E

- The LUCIA tool enables the exchange of cyber incidents, allowing its adaptation to other ticketing tools used in the organisations, improving the Request Tracker for Incident Response tool (RT-IR) and making it multi-agency with the capacity to federate between the tools used so that the exchange is fluid. Finally, it provides efficiency metrics concerning the resolution of cyber incidents;

- The REYES cyber intelligence tool attempts to tailor this information to the needs of the agencies by providing them with its exposure surface (what the attacker knows about it, vulnerabilities, compromised passwords and other valuable information). Cyber intelligence is also distributed to improve perimeter protection. Finally, the tool provides the necessary information on unknown IPs and domains.

The Royal Decree 421/2004, of March 12th, which regulates the National Cryptologic Centre, establishes the scope of action of CCN as the "security of the Administration's information technology systems that process, store or transmit information in electronic format, that require protection by law, and that include means of encryption" and "the security of information technology systems that process, store or transmit classified information". Considering the current geopolitical situation, how do you assess the probability of the occurrence of a severe attack targeting the systems under the scope of action of CCN?

The probability of a cyber-attack occurring is very difficult to establish. Systems must be prepared in the areas of PREVENTION, DETECTION and RESPONSE.

CCN-CERT promotes actions in these three fields to improve the efficiency of public bodies. In order to do this, the RD 421/2004 is supplemented by the RD 311/2022 of the National Security Framework (which is an update of RD 3/2010). The functions of CCN-CERT are extended here:

- It is the National Government CERT;

- The rapid cyber-attack response enables deployment of early warning systems;

- It provides alerts on vulnerabilities in the technology and within the agencies themselves (due to configuration deficiencies in their systems);

- • Deployment of rapid response teams (RRTS) to critical incidents.

Since the beginning of the war in Ukraine, are you aware of incidents in the physical infrastructure (terrestrial cables, submarines) that make cyberspace possible in the Kingdom of Spain or cyberattacks directed at the systems of the general state administration and its public bodies with the purpose of extracting data? What kind of cooperation in this regard have you established with other CSIRTs/CERTs of international partners and allies?

No major incidents directly related to the Ukrainian war have been detected. Statesponsored cyber-espionage actions against sectors of interest to adversaries in the Ukrainian war have been detected.

Theft of sensitive information, intellectual property or state secrets is a common occurrence in our government networks. At least 20 have been detected during the year. No more were detected in 2022 than in previous years.

CCN-CERT has many alliances for the exchange of information, including the European Government CERT (EGC), an informal group of government CERTs in which a lot of valuable information is exchanged.

How does the CCN develop its mission and function in the framework of Spanish foreign policy?

CCN-CERT carries out many training exercises, dissemination of best practices, setting up of SOC/CERT and assistance in the response to critical incidents in many Latin American countries. In addition, the external service is subject to special protection by CCN-CERT.

How many cyber-attacks targeting the public administration systems happen in Spain each year and what percentage of them are related to activity by hostile foreign intelligence services or intelligence activities by non-state actors aiming to exfiltrate classified information or harm national security?

CCN-CERT classifies cyber incidents into five danger levels: LOW, MEDIUM, HIGH, VERY HIGH and CRITICAL. We especially monitor VERY HIGH and CRITICAL incidents.

In 2022, about 55,000 incidents affecting the public sector have been managed. Of these, approximately 70 were recorded as CRITICAL. Of these incidents, 60% are related to cybercrime groups (ransomware) or theft of personal information. The remaining 40% are related to state-sponsored groups (about 30 cyber incidents).

In the latter incidents, if detection is late, it is difficult to determine the volume and type of information extracted, so the impact is difficult to gauge.

— What are the main current state-based threats targeting the cybersecurity of Spain? Could you describe some examples of incidents during 2022 involving states actors or APTs?

State-sponsored attacks have focused on the theft of sensitive information from public bodies and companies (in this type of attack, CCN-CERT has the responsibility to act and help the body determine the scope of the attack and remove the threat from its networks).

APT-related incidents are usually classified and it is not possible to report on their scope and typology.

— One of the main challenges of hostile activities in the cyber domain is early detection and attribution, could you explain how you gather evidence and assess threat actors involved in cyber-attacks?

CCN is part of the Spanish Intelligence Service (CNI). The attribution is carried out with cyber intelligence units that use information provided by CCN-CERT and other CNI capabilities.

A lot of information about attackers' tactics, techniques and procedures (TTPS) is exchanged to determine attribution with a high degree of certainty.

— Have you observed a surge of cyber-attacks during the months preceding the aggression against Ukraine and during the NATO Summit in Madrid?

We did not observe any escalation in cyber-attacks detected during the NATO Summit.

— What is the frequency of cyber-attacks targeting critical infrastructure in Spain?

In Spain, critical infrastructure is subdivided into 12 sectors.

CCN-CERT only has visibility on critical infrastructure corresponding to the public sector (government, health, water, transport maritime, rail or underground, food, research facilities).

The detection of cyber-incidents has the same parameters as the general distribution.

What emerging and current trends in the cyber and other domains are deemed to have a significant impact on cybersecurity?

In the last two years the biggest impact on cyber security has been:

- The use of public cloud services as a complement to the services provided by the corporate network;
- The extension of remote work in the organisation's activity. The cybersecurity measures have often not been increased to protect this new working model;
- The use of corporate or personal mobile devices receiving sensitive information from the organisation without adequate protection measures;
- The professionalisation of strike groups.

All these new paradigms make it necessary to work in the fields of:

- PREVENTION: reduce the exposure surface;
- DETECTION: continuous 24×7 surveillance;
- RESPONSE: by requiring cybersecurity operations centres that provide this capability horizontally to a variety of agencies.

Could you explain how CCN-CERT conducts early warning procedures and how the Early Warning System (SAT) operates?

The early warning systems operated by CCN-CERT began to be developed in 2008. They are deployed on a voluntary basis in agencies. They are rule-based systems that provide detection capabilities for known cyber-attacks. The detection rules are updated daily and use both their own and external sources (commercial feeds).

The logs generated can be exploited by the organisation or alerts can be received directly from CCN-CERT.

And they cover the following fields:

- SAT SARA. It provides detection capacity to the Governmental Network (RED SARA) that interconnects the central administration, regional governments and local entities. An aggregation of logs from the perimeter devices of the connection areas of the different bodies that are connected is carried out. There are 50 sensors;

- SAT INTERNET. Intrusion detection system in the Internet traffic of the organisations. It includes various technologies to fine-tune this detection capability. There are more than 320 sensors available;

- SAT ICS. It is a detection system for industrial networks or corporate networks that include many devices using industrial protocols (e.g., hospitals). It performs asset survey and dissection of industrial protocols. More than 50 sensors are available.

These systems can detect known attacks and enable rapid defence against new cyber-attacks or new malware samples. All detection is based on the analysis of network traffic.

The CCN website is remarkable for the information and public reports shared. Could you explain what other kind of information products (non-public) you produce?

Regarding the public reports, in addition to the CCN-STIC guides we can point out the threat reports (IA), malicious code analysis reports (ID) and best practice reports (BP). Most of these reports are available on the website.

Furthermore, CCN-CERT prepares Technical (IT) reports that are associated with the investigation of cyber-incidents or the performance of audits. These reports are sent to the body or bodies concerned. In 2022, around 100 reports were produced.

What measures of the National Cybersecurity Plan, approved by the Council of Ministers on March 29, 2022, did you consider most relevant and are having or have the potential to make a greater impact?

The measures that will have the greatest impact will be those related to:

- Boosting the implementation of the ENS;

- Development of Cybersecurity Operations Centres and their integration into the National SOC Network;
- Development of active cyber-defence measures to take the initiative against cyber-attackers;
- Development of training/awareness raising systems to improve their level. These actions should be accompanied by corresponding metrics.

How does CCN-CERT work in the prevention, detection, and neutralisation of cyberthreats and operations in cyberspace by threat actors?

Cyber operations on cyberthreats are offensive activities that are not the responsibility of CCN-CERT.

Does the CCN and/or CCN-CERT have a role against foreign information manipulations and interference within cyberspace? Do you monitor hostile narratives and disinformation from foreign state actors targeting Spain?

The disinformation activities are not the responsibility of CCN-CERT.

What capabilities or measures has Spain developed against emerging threats like Deepfakes or synthetic content for cyber and influence operations by foreign and criminal actors?

This activity is not the responsibility of CCN-CERT.

From the perspective of training curricula, what gaps/needs in competences or skills have you identified, or do you consider to be very relevant for strengthening cybersecurity?

We need very specific technical skills in system audits, investigation of cyber incidents (in particular, reverse engineering, mobile phone analysis and mass LOG analysis).

What relative percentage of cyber incidents in Spain originate from domestic vs foreign threat actors?

Most cyber incidents (80%) originate from external actors.

Do you conduct simulations or wargames with stakeholders in the scope of action of CCN to assess potential cyber or crisis scenarios or test gaps and needs in capabilities?

We conduct cyber exercises focused on crisis management of cyber-attacks by statesponsored actors (APTs) and cybercrime groups (ransomware).

They are based on the CCN-CERT's experience in dealing with this type of attack.

How do you foresee the future of cybersecurity? What technologies have a greater disruptive potential?

Artificial Intelligence based on detected traffic will allow analysts to focus on tailored cyber-attacks. All of our systems must migrate to AI-based technologies.

This discipline should be complemented by threat-hunting activities based on the experience of analysts.

On a scale of 0 to 10, how do you assess the state of cybersecurity culture in Spain? Do you think the autonomous communities and local administration have a high level of commitment to cybersecurity?

Awareness has increased considerably in Autonomous Communities and Local Government following the ransomware attacks we have undergone. We estimate that we have moved to a level of at least 8. The role of cybersecurity managers has been greatly strengthened.

What specific measures could promote and bootstrap SME's cybersecurity or their cyber hygiene practices?

In Spain we have an adaptation of the National Security Framework for small local administrations with 35 adapted measures that allow for adequate cybersecurity. This system could be adapted to small and medium-sized enterprises also.

2023 will be an election year in Spain – what initiatives has Spain developed against election interference and for countering potential malicious activities in the Spanish civilian cyberspace in this context?

CCN-CERT carries out a deployment during elections that supports the Ministry of Home Affairs or the corresponding Autonomous Community to protect the electoral system from cyber-attacks.

Activities include:

- Audits of all systems involved, identifying and prioritising the vulnerabilities to be addressed;
- Cyber-surveillance campaign to identify possible actors that could carry out attacks on deployed systems;
- Continuous surveillance during Election Day until the presentation of results to the public.

Other comments or aspects that you consider relevant from the CCN's point of view regarding the Kingdom of Spain's cyberspace and cybersecurity?

The protection of Spanish cyberspace is the CCN's fundamental objective. To achieve this, it is necessary to count on all the agents that contribute to this activity and to form a SINGLE CYBERSHIELD as the motto of our XVI Jornadas STIC CCN-CERT, the largest cybersecurity event organised in Spain and held just a few weeks ago.

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Examining Supply Chain Risks in Autonomous Weapon Systems and Artificial Intelligence

Austin Wyatt | RAND Australia, ORCID: 0000-0003-1901-8019

Abstract

The development of increasingly AI-enabled autonomous systems and other military applications of Artificial Intelligence (AI) have been recognised as emergent major military innovations. In the absence of an effective and enforceable ban on their development and/or usage arising from the Group of Governmental Experts on Lethal Autonomous Weapon Systems (LAWS), it is likely that such systems will continue to be development. Amongst the legal, ethical, practical, and strategic concerns raised by the emergence of such systems, it is important not to lose sight of the risks involved in relying on a high-manufactured system in place of a human. This places additional strains and importance on securing diverse, complex, and over cross-jurisdictional supply chains. This article focuses on the vulnerability of and the risks to the integrity and security of the supply chains responsible for producing AI-enabled autonomous military systems.

Keywords

supply chain risk, autonomous weapon systems, Artificial Intelligence, emerging technology

Received: 7.06.2023

Accepted: 17.11.2023

Published: 29.11.2023

Cite this article as:
A. Wyatt "Examining Supply Chain Risks in Autonomous weapon systems and Artificial Intelligence," ACIG, vol. 2, no. 1, 2023, DOI: 10.60097/ACIG/162874

Corresponding author:
Austin Wyatt, RAND Australia; ORCID: 0000-0003-1901-8019; E-MAIL: awyatt@rand.org

1. Introduction

The increasing and continuous development of autonomous systems and other military applications of that consist of and/or include Artificial Intelligence (AI) have been recognised as an emergent major military innovation [1]. This recognition is underscored in strategic documentation from both China [2] and the United States [1], along with attendant large-scale investments from both state and commercial actors. Recognition of AI as an emergent major military innovation in the US and China is particularly important because they are locked in hegemonic competition in the Asia Pacific and account for the vast majority of military expenditure, both in terms of procurement and research and development. Amongst the benefits advanced in support of developing AI-enabled autonomous systems, their potential to safeguard soldiers by removing them from the direct line of fire is commonly cited. Another benefit is that AI-enabled systems confer a strategic advantage by facilitating tactical and operational decision making at a pace exceeding human capability. Even if one assumes that these benefits are achieved, such systems also raise numerous ethical and legal issues [1], as well as arguably increasing risk to non-combatants [3] – commonly referred to as collateral damage. Whether autonomous systems will be a net positive or negative influence on the future of warfare ethics, in the absence of significant advances in the efforts at the Convention on Certain Conventional Weapons [1] toward an international legal ban,[1] they are likely to remain prominent in the future paradigm of conflict, and thus have significant impact on the achievement national security objectives.

It is therefore important that the international community considers the viability and attendant risks involved in relying on military applications of AI, including AI-enabled autonomous weapon systems. AI is best thought of as an enabling innovation, closer to electricity than the machine gun [50], and is the core component of any autonomous systems. Part of the challenge in doing so is the lack of a universally agreed set of terminology for engaging in the debate around responsible use of AI in the military domain[2], (beyond that they are generally non-deterministic complex systems without an integrated human operator. Even lethal autonomous weapon systems, arguably the most problematic AI-enabled military technology category, lack a universal definition [1]. While most writers start with the definition presented by US Directive 3000.09 [4][3], other definitions abound including from the UK Ministry of Defence [5], and the Australian Defence Force [6], as well as from academia include those published by Horowitz [7], Scharre [8], Roff [9], Wyatt [1], and Bode, Huelss and Nadibaidze [10].

1 — The United Nations Convention on Certain Conventional Weapons (CCW) has held semi-annual meetings of selected governmental experts on Lethal Autonomous Weapon Systems since 2014. These meetings were part of an effort to develop international legal instruments for governing the use of AI-enabled autonomous weapon systems and was the main arena in which a pre-emptive ban was mooted. Although its consensus-based approach has not yet yielded conclusive results, it did support the passing of the first General Assembly resolution on autonomous weapons in 2023.

2 — This refers to applications of Artificial Intelligence and autonomous systems within the context of military operations. Adopting this lens restricts the discussion to the incorporation of AI and autonomous systems into the generation and employment of military power, ranging from logistics to weapon systems, but excludes purely commercial applications. This term was prominently used by the Netherlands, hosts of Responsible Employment of Artificial Intelligence in the Military domain (REAIM 2023) Conference.

In the absence of a universally agreed definition for AI-enabled autonomous systems, it has become increasingly commonplace to leverage the three functional autonomy categories developed by the International Committee of the Red Cross and Human Rights Watch. This approach categorises systems are categorised by their ability to execute its 'critical functions' independent of a human operator [11]. An AI-enabled system could, therefore, be described as a supervised autonomous systems (where a human remains on the decision loop, where they can interrupt the system's actions), semi-autonomous systems (where the system has a limited capability to act autonomously within geographic or functional limitations, and a human remains in the decision loop), and fully autonomous systems (where the system has effectively independent control of its critical functions, removing the human from the decision loop) [1].

The absence of a human operator places even more emphasis on the reliability and effectiveness of the system itself. A sufficient level of certainty and safety inherent in these systems is not merely contingent on the technology maturing to a set future point [12]. Instead, one must take a more holistic approach, one that considers other elements and actors that may influence or compromise the effectiveness and reliability of future AI-enabled autonomous weapon systems, whether due to error or malicious action.

This article focuses on exploring the risks associated with the integrity and security of the supply chains responsible for producing AI-enabled autonomous systems. Supply chains are, by their nature, complex networks with multiple nodes and links, each vulnerable to potential disruptions and security breaches [13]. Such networks typically span geographic and jurisdictional boundaries and are reliant on many of the same key transit points as more general global trade. The dislocated nature of the global supply chain for AI-related technologies and the wide range of civilian as well as military actors increases the complexity of securing accountability [56]. Disruptions to critical technology supply chains, such as those associated with the military industrial complex and associated national security operations, could delay or prevent the deployment and maintenance of AI-enabled autonomous systems during times of increased competition or conflict. Fedasiuk et. al. highlight the potential for an adversary to hamper western access to crucial advanced chip sets [57], a risk that is particularly concerning given the vulnerability of the main supplier of such chips, Taiwan, to China. Uniquely to AI-enabled systems, there are also risks involved in the supply chains for the data that make these systems intelligent. Morgan et. al. suggest that even air-gapped AI-enabled systems, while more resilient against hacking,

remain vulnerable to degradation through data poisoning attacks, where an adversary maliciously injects code into the training data to fool the resultant system, or via physical adversarial attacks, such as specially designed stickers that fool computer vision algorithms [58]. Resilience must be built early into supply chains to ensure that such systems are not compromised by contamination of their training data or the insertion of zero-day exploits [14][4].

Given the breadth of AI-related risks in logistics, this article limits itself to exploring the supply chain risks that could stem from adopting AI and AI-enabled autonomous systems. This paper is also intended to provide a broad introduction to the issues, a further exploration of these issues from the perspective of particular military or regional perspective.

2. AI, Autonomous Systems and Future of Conflict

The rise of AI-enabled Autonomous Weapon Systems (AWS) as a potential Revolution in Military Affairs (RMA) is anticipated to have a revolutionary, and thus as disruptive impact on the future of conflict. Despite popular belief, innovation requires both the maturation of an invention and the development of operational concepts to utilise that invention in a disruptive manner [15]. This does not merely represent a pioneering deployment of an autonomous system by a state; instead, different states might opt for unique development strategies for related technologies or pair a matured autonomous system with distinct yet non-revolutionary operational concepts [15]. Moreover, developers may adopt strategies to limit the exposure of their methods to safeguard operational advantage or to avoid international scrutiny, particularly in the case of LAWS [16]. Eventually, a state will introduce a fully autonomous weapon system that disrupts conventional military balances,[5] compelling other states to react to the resultant shift in relative power [17] or relative advantage. This pivotal moment, known as the demonstration point, obliges competitor states to respond or concede strategic advantage to the initial state deploying such systems [15].

However, first mover advantage may be transient, particularly when it pertains to disruptive innovations like AI-enabled autonomous systems, hypothesized to possess low proliferation barriers [1]. Past military innovations typically demanded considerable resources or organizational capital, limiting the ability of states to respond to a demonstration point by matching the initial mover's advancements.

4 — Zero day exploits are vulnerabilities in a computer system that are unknown to the users or manufacturers until they are deployed by an adversary. These are particularly concerning in the case of AI-enabled systems because of the complex nature of such systems. This risk is further exacerbated by the vulnerability of AI to corruption (whether deliberate or not) of the underlying training data set.

5 — This refers to the relative capacity of states to "adopt the key military methods of a period" [15], which in the current paradigm could include precision munitions, space-based communications, and aircraft carriers.

When faced with novel forms of warfare, for example, the advent of aircraft carrier warfare, a less powerful state theoretically could attempt to surpass the first mover, but practical constraints of resources and political will would limit this [15]. The adaptation/adoption of doctrine, not a trivial matter, also needs consideration. Yet, if the barriers for entry and early adoption are significantly diminished (due to the dual-use nature of related technologies or lack of need for specialized skill sets), the disruptive effects of rapid proliferation to multiple state and potentially non-state actors should be considered [18], as exemplified by the widespread use of remotely operated armed drones [19]. Such proliferation would also have a significant impact on supply chain vulnerability once these technologies become widely distributed.

Predicting the precise effects of such proliferation on future warfare remains challenging. However, historical insights from military and civilian disruptive innovation theory, such as aircraft carrier warfare [15], coupled with the unique attributes of LAWS (and other AI-enabled military technologies), as well as initial state reactions to their early development, provide a first-order, yet robust foundation for hypothesizing potential outcomes. An overarching characteristic of major military innovations is their transformative influence on how states project power and conduct warfare [1]. Historically, this has precipitated disruptions in the international balance of power, providing opportunities for middle and minor power states to challenge existing hegemonic power balances, in both global and regional contexts. This change can enable a rising challenger state (such as China) to counterbalance the traditional advantage enjoyed by the existing hegemon (in this case the US), while smaller states strive to mimic successful states (e.g. Taiwan mimicking the US) to safeguard their own power bases from their rivals, thereby accelerating diffusion [20]. Threatened by the deterioration of its relative advantage, the incumbent state is induced to adopt or enhance the tempo of its Revolution in Military Affairs (RMA) efforts to regain its standing [1]. This diffusion of major military innovation may engender regional instability and precipitate hegemonic warfare [21] – typically referred to an arms race or a negative security cycle within the realist paradigm. Given the relatively low adoption barriers for autonomous weapon systems compared to prior major military innovations like nuclear weapons, and the comparative difficulty in applying conventional arms control mechanisms [51], it is hypothesized that the emergence of LAWS will have a destabilizing influence on the future of warfare.

From a grand strategic standpoint, the potential for middle and minor powers to emerge as successful early adopters of AI-enabled

autonomous systems represents a departure from historical precedents, like nuclear weapons, where middle and minor powers were compelled to align with a great power competitor to protect their interests [2]. Instead, states in the global South could potentially exercise greater autonomy, balancing competing great powers regionally while deterring aggression from similarly sized neighbours. This could instigate an escalating cycle of arms acquisition and posturing as regional powers deploy systems lacking effective legal or normative controls, thus intensifying security dilemmas [20]. Without mutually accepted norms around appropriate uses and responses to such systems or effective international legal treaty banning their use (for example through the CCW), there is a considerable risk of unanticipated escalation, whether between the great powers or between regional powers in Southeast Asia or Africa for example. Additionally, the spread of remotely operated, autonomous, and/or AI-enabled systems, especially given the dual-use nature of enabling technologies, poses a significant risk of these systems falling into the hands of violent non-state actors. The result may be a less stable balance of power, particularly in the Asia-Pacific, leading to a multipolar military competition domain rather than a traditional hegemonic transition of power.

While considering the influence of these systems on regional stability and the likelihood of new conflict or the prolonging of existing conflict, it is important to debunk two persistent myths surrounding AI-enabled autonomous weapon systems. The first is the fear of a 'Terminator' being developed or deployed in the foreseeable future.[6] Designers and potential state end-users are rational actors who are generally cognizant of the ethical issues raised by LAWS.[7] Admittedly, this would not apply to violent non-state actors such as terrorist groups or extremist individuals. Secondly, the rise of AI-enabled autonomous weapon systems does not signify that future wars will become 'bloodless' or 'sterile' [25]. War remains a human endeavour, and human casualties, particularly among civilian populations in urban operations, are unfortunately inevitable – be it intentional or collateral. Both of these perspectives oversimplify the issue, disregarding the more plausible scenario of widespread deployment of these systems disproportionately affecting the technologically inferior adversaries [26]. The introduction of autonomous systems raises significant ethical challenges, particularly regarding the kill-decisions [3]. Simultaneously, there is a moral obligation on leaders to utilize autonomous systems where they can protect the lives of soldiers, even if their deployment is limited to the dull, dirty, and dangerous roles [27]. With all this said, it would now be pertinent to consider the vulnerability of supply chains as single points of failure for the security of these systems.

6 —— For example the discussion by Shead and the concept of 'slaughterbots' [22].

7 —— While this is a broad claim, it is supported by perception studies focused on Machine Learning developers [23] and ADF personnel [24], as well as the recent call from Open AI's CEO for greater regulation of the area.

3. Supply Chain Risks

Increased reliance on AI-enabled systems also increases the variety and seriousness of vulnerabilities in the supply chain. AI-enabled Autonomous systems would not be reliant on a human for critical functions [10]. In addition to the myriad legal and ethical challenges this change poses, however, it also quite simply places the entirety of the burden for that system to run effectively, reliably, and safely on manufactured components. There is no human to recognise and correct errors; for example, that the scope of a rifle was incorrectly zeroed or that a civilian aircraft has been mis-identified as a legitimate target. Ensuring that AI-enabled systems operate as expected and fail safely thus become crucial characteristics, yet they are dependent on securing disparate and often complex trans-regional and trans-national supply chains. In the following section, key geo-strategic-, technological-, and economic risks to these chains will be examined.

3.1. Geostrategic Supply Chain Risks

Beginning with the geostrategic risks associated with the supply chains for AI-enabled systems sensibly reflects the recognition of the likely importance of such systems to the future of warfare. Further, even amongst states that do not see themselves as a potential first mover, the strong public commitment to AI and autonomy by the US and China encourage smaller states and violent non-state actors to invest in mechanisms for countering the advantage offered by such systems, with the supply chain being a novel and comparatively vulnerable attack surface.

First, despite the recent surge in public accessibility of Large Language Models, machine-learning based complex AI remains expensive [28] and reliant on large amounts of computing power [29], cooling [30], and above all, data (which raises its own ethical and legal challenges) [31]. Reliance on a global supply chain diminishes the capacity of states from a sovereign control perspective, particularly non-great powers, to guarantee access and to impose sufficient security controls over the manufacturing and development process. For example, rare earth metals, crucial for many advanced technologies, are primarily sourced from a few countries, presenting a geopolitical risk if these countries decide to leverage their monopolistic control over these resources. In the event of geopolitical tensions, a trade war or even an embargo, their access to critical resources may be limited. A case in point would be a blockade of Taiwan could have immediate and disastrous effects on high-technology supply chains internationally [32]. The current sanction regime against Russia due

to the illegal war with Ukraine provide ample real-time examples of how military industrial complex and dual-use supply chains affect the ability of even a superpower to maintain (relative) advantage.[8]

Relatedly, and of particular concern for states such as Australia, international technological controls and regulations can impact the availability and transfer of technology, particularly for emergent or particularly sensitive systems. International Traffic in Arms Regulations (ITAR) is a particularly well-known example of how countries may restrict the export of technologies deemed either critical to national security or related to maintaining a particular capability advantage [33]. Additionally, international regulatory bodies – for example the Wasserman Arrangement – of which Australia is a signatory nation amongst 44 other nations[9] – may impose restrictions or sanctions on AI and autonomous system-related technologies or developers. For example, the international community continues to debate whether a ban on autonomous weapon systems is appropriate, or likely to be effective. Supply of critical components could be limited or blocked if such a ban were implemented, or if individual states or a bloc – such as the EU – were to implement their own restrictions. This risk would be particularly troubling if AI or autonomous weapon systems come to rely on a single source for a critical component, such as high-capability semi-conductors produced primarily in Taiwan. Such dependency creates a strategic vulnerability because any disruption to the supply from this source – due to logistical issues, manufacturing constraints, or other factors – can result in severe manufacturing and subsequent operational setbacks – severely affecting the ability to secure national interests. It also gives the supplier considerable leverage, potentially leading to increased prices, unfavourable terms and/or even insisting on being included into economic/defence pacts such as NATO or the Five Eyes Alliance.

Contrastingly, a diversified, multi-jurisdictional supply chain increases the risk of intellectual property (IP) theft or duplication, as well as the potential for proliferation of such systems to smaller states and violent non-state actors. The development and application of AI in military contexts often involve proprietary algorithms, data models, and technologies, representing substantial intellectual capital. This sensitive information, if leaked or stolen, could significantly undermine a nation's technological edge and compromise its national security. Supply chains that span across multiple countries and vendors increase the risk of such IP being compromised, especially if these entities have differing or inadequate cybersecurity measures and different levels of security consciousness. Consequently, it becomes

8 — The author would like to thank Dr Dries Putter for this example.

9 — The author would like to thank Dr Dries Putter for suggesting the Wasserman Arrangement as an example.

crucial to ensure robust protection of IP across the entirety of the supply chain [13] which will require significant counterintelligence measures and thus increasing the unit costs concomitantly. In the absence of such protections, there is also a risk of uncontrolled proliferation, exacerbated by the dual-use nature of the underlying technologies. The risk of such systems falling into the hands of adversarial or rogue states, non-state actors, or even terrorist groups is a significant security risk that's mitigation is complicated by cross-jurisdictional supply chains involving multiple civilian actors. This technology proliferation can lead to an advantage leveling effect on the strategic landscape, increased risk and severity of extremist/terrorist violence, and could raise the security vulnerability for states that would not otherwise vigorously pursue such weapons.

Finally, given the importance publicly placed on AI by leading militaries (such as the United States, China, and Russia), one must also consider the risk of a malicious non-state actor (whether a disgruntled insider threat, terrorist group or extremist) deliberately interfering with, disrupting, delaying or degrading critical supply chains, reducing or eliminating the capacity of a state to produce and maintain key military systems [34]. The principal risk surfaces for this are arguably in the cyber domain, particularly in the face of US decoupling and friend-shoring efforts. A key mechanism for this sort of malicious action is through the cyber domain. The low entry cost of operations in the cyber domain (whether attack, subversion, or intrusion) encourages their use by smaller actors, particularly those utilising existing AI tools as force-multipliers. For such actors, the opportunity to disrupt or sabotage high-capability high-cost systems through exploiting vulnerabilities in their supply chains, is an attractive levelling mechanism. Such attacks range from the theft of critical intellectual property [13] to the malicious manipulation of training data [35]. The high level of interconnectivity in global supply chains, as demonstrated by Covid-19, and the widespread use of digital systems in both official tasks and in the homes of related individuals exacerbate these risks [36]. Even onshoring sufficient manufacturing capabilities to produce key components domestically does not eliminate these risks, as interdiction could be launched lower down in the supply chain, at the raw materials level, for example [34].

3.2. Technological Supply Chain Risks

The first technological supply chain risk pertains to technology obsolescence. Given the rapidity of technological development in the field of AI, there is a substantial risk that any procured technology may become obsolete soon after acquisition. This fast-paced

evolution is fuelled by constant advancements in algorithms, data processing capabilities, and computational hardware – fuelling the RMA concept internationally. The implications of technological obsolescence are multifaceted and significant. First, the financial resources invested into the design, manufacturing, acquisition, integration, and training of personnel for specific AI technologies could become sunk costs if these technologies rapidly become outdated. This risk is exacerbated by the typically protracted defence procurement processes, which often lag behind the pace of technological advancements. The discrepancy in pace between procurement and technological progress could result in the acquisition of technologies that are already verging on obsolescence at the point of implementation. Operationally, the consequences could be equally detrimental. Outdated AI technologies could impair a military force's effectiveness, potentially leading to tactical and strategic disadvantages in the field. Moreover, support for older technologies may diminish as manufacturers and software developers move towards more advanced and efficient models, making it difficult to maintain and repair existing systems. Lastly, as AI technologies continue to evolve and proliferate globally, maintaining up-to-date systems is paramount as operating outdated systems could expose vulnerabilities to potential adversaries and compromise the security and effectiveness of military operations and thus national security. The need to avoid obsolescence in not only the end product but the key production nodes for such systems makes it imperative for militaries and manufactures to adopt an agile approach to technology acquisition and implementation. This could involve shortening procurement cycles, investing in regular technology refreshment programs, and establishing collaborative partnerships with technology providers to ensure early access to cutting-edge AI technologies [37]. Additionally, incorporating flexibility in procurement contracts to accommodate technological upgrades can also help in keeping pace with rapid advancements [38]. Of course, it has to be said that increasing the tempo of technology uptake in organisations will also open the vulnerabilities to increased levels of corruption and graft typically associated with defence contracting – and hence the long and bureaucratic procurement processes to ensure transparency and accountability. Thus, corruption due to the requirements for agility within supply chains poses another distinct challenge to security.[10]

The second major risk involves the complexity and fragmentation of supply chains inherent in the production and deployment of AI technologies. These supply chains often stretch across the globe, involving various suppliers for essential hardware components,

10 —— The author would like to thank Dr Dries Putter for this point.

software applications, and data resources. This complexity and fragmentation engender a multitude of risks. For one, a disruption at any point in the supply chain, whether it's a failure to produce a critical hardware component, a disruption in logistical operations, or a software development issue, can have a significant downstream effect. This can potentially delay or even halt the delivery and deployment of AI technologies, severely affecting the military's operational readiness and capabilities. The fragmentation of the supply chain also raises issues regarding quality control and security. With multiple suppliers involved in the production process, maintaining consistent quality standards across all components becomes challenging. This was illustrated repeatedly with the security challenges faced by the F-35 development and production efforts [39]. Similarly, with so many points of entry in the supply chain, the risk of malicious actors introducing vulnerabilities into the system is significant. Mitigating these challenges could take the form of friend-shoring, supporting the development of alternative suppliers of key components and raw materials in allied states in order to reduce the threat surface [40], or implementing (contractually or through legislative tools) strong quality control and cybersecurity protocols across key nodes of the supply chain. However, these issues aside, there is also good reason for having fragmented supply chains – i.e. fragmented insight into the total composition of a sensitive system. Thus, there needs to be a balance between the requirements for supply chain fragmentation and the need to security.

The third technological risk relates to the potential vulnerabilities of AI systems themselves. These could be due to design flaws, manufacturing defects, malicious interference, or software code malfunctions [14]. Unfortunately, due to the tendency of such complex systems to fail, the results of such vulnerabilities in a military context could be severe, and would damage vital trust between the system and its human user/supervisors even if the malfunction does not cause physical harm. Such trust is a integral part of technology adoption by the organisation to the point where doctrine is written for it or adapted to accommodate it. This risk is exacerbated by the opaqueness nature of certain AI systems, the so called black-box problem [41]. The lack of transparency, explainability, and common understanding of an AI-system's functionality makes it difficult to predict and understand system behaviour, especially in emergent situations [42], for example the recently disclosed thought experiment in which an AI system operating in a simulated environment eliminated its (simulated) human overseer in order to maximise its capacity to fulfil its core mission [52].

Cybersecurity threats represent a further significant concern. As digital systems, AI technologies are attractive targets for cyberattacks that could degrade or disable them in store or on the battlefield [14]. The source of these cyber threats could range from state-sponsored actors aiming to disrupt military capability, to violent non-state actors such as terrorist groups or organized crime syndicates seeking to exploit vulnerabilities for their own ends. Importantly, these vulnerabilities could be introduced at any stage of the supply chain, underscoring the necessity of end-to-end security measures – with concomitant cost implications for the end user.

3.3. Economic Supply Chain Risks

Another salient aspect in the discourse on the adoption of AI by militaries is the substantial expenditure associated with the development, deployment, and maintenance of these advanced systems. It is an inherently resource-intensive pursuit, requiring considerable investment in various facets of the development and procurement processes. The research and development phase, the cornerstone of AI evolution, demands a prodigious financial commitment and human capital outlay [43]. Furthermore, recruiting and retaining skilled personnel capable of undertaking such complex development tasks also represents a significant financial undertaking, especially considering the high demand for these experts in the competitive technology market [44] and the further cost of security vetting and maintaining security from an insider threat perspective – the Edward Snowden incident being a case in point. Access to the best-quality (highly qualified and/or experienced doctoral qualified researchers) specialist talent, including data scientists and machine learning experts, is pivotal to driving innovation and improving the operational efficiency and reliability of military AI applications [12]. For example, acquiring top-level talent is a known barrier in Chinese military AI efforts, due to government policies and the opportunities offered by working in the US or Europe [2]. Once an AI system is developed, the procurement process entails substantial funding [44]. The hardware, comprising high-speed processors and robust storage solutions, forms the backbone of AI capabilities. Simultaneously, software and data acquisition are critical for the system's decision-making ability, as it feeds and trains the underlying algorithms. Additionally, AI technologies require ongoing updates and maintenance, further escalating the overall costs. This continued investment is essential to keep abreast of rapid technological advancements, ensure system security, and mitigate potential obsolescence. These updates may encompass software patches to enhance the system's capabilities or address vulnerabilities, hardware upgrades to improve performance,

and data management activities to ensure integrity and compliance with regulations. Given the high costs involved in being the first mover for AI systems, reliance on AI in military applications poses potential risks to defence budgets. There is a plausible concern that the financial burden of attempting to maintain a capability offset based solely on technological advantaged derived from advanced AI capabilities may strain defence budgets. Such arguments must be balanced against the argument that autonomous systems and other uses of military AI offer significant potential dividends in terms of enhanced operational efficiency, precision in decision-making, and maintenance costs over their life of type. While initial costs are generally exorbitant (and hence also resulting in very high entry barriers for competitors), AI systems have proven to be far cheaper to duplicate and diffuse once in use, meaning that the bar to entry for fast followers is significantly less resource intensive than this section would initially suggest [45]. Again, emphasising the enormous responsibility of national counterintelligence capabilities to secure the IP at every point in the supply chain to ensure entry barriers remain high and threat actors are barred from access. Thus, an escalation in cost.

However, the initial costings for a first mover can also prove unpredictable. Developing autonomous systems involves the procurement of sophisticated hardware and advanced software, along with the accumulation and management of vast amounts of data. These elements are essential to construct, operate, and regularly update the system. However, these components can be susceptible to considerable price fluctuations. The unpredictability of costs is largely determined by changes in market supply and demand, escalating geopolitical tensions, or dramatic shifts in economic policies. These factors can significantly alter the costs associated with the development, operation, and maintenance of military AI. Moreover, these uncertainties can hinder strategic planning and the ability to forecast future requirements accurately. They also complicate the allocation of defence budgets, which are typically subjected to rigorous scrutiny and oversight. For these reasons, the inherent cost volatility and unpredictability represent one of the most significant risks in integrating AI into military systems, especially because of the requirement for public scrutiny and accountability for projected spending vs. the value proposition. The development, implementation, and maintenance of AI-based systems necessitate a wide array of resources. This ranges from physical materials such as rare earth metals essential for manufacturing advanced electronics (chiefly Lithium [53]), semiconductors critical for data processing, and extensive data storage infrastructure, to human resources

with specialized skills in AI research, development, and deployment. A scarcity or interruption in the availability of any of these critical resources may lead to significant supply chain disruptions and costs [44]. Similarly, the known challenges faced by militaries to recruit and retain experts in relevant fields poses a significant barrier to the widespread adoption of AI in the military [44]. This deficit is further exacerbated by the fierce competition for talent from the private sector, where compensation often exceeds what the public sector can offer. Therefore, any strategic plan for the adoption of AI in the military must include a robust strategy for resource acquisition, management and retention to mitigate these risks.

An often overlooked yet significant factor in the discussion about the implementation of AI in military applications (such as autonomous aircraft) is the economic cost associated with regulatory compliance. Adherence to both international norms and domestic regulations governing the use of AI can impose substantial costs on defence departments and the associated industries. To start with, one of the significant expenses associated with regulatory compliance is related to data protection and privacy. The use of AI technologies often involves the processing of vast amounts of data, some of which may be personal or sensitive. Complying with data protection regulations can necessitate significant investments in secure data storage and processing infrastructure. For example, compliance with Europe's General Data Protection Regulation requirements was expected to cost large business an average of 1.3 million euros [54]. It also involves the continuous updating of security protocols and measures to prevent unauthorized access or data breaches. Implementing robust data management policies and procedures that are compliant with privacy laws, which vary from nation to nation, is a complex and costly task, but it is essential given the sensitive nature of military operations and the potential for misuse of personal data. In addition, certification processes can be costly and time-consuming, but they are often a necessary step in demonstrating that a system is safe, reliable, and compliant with regulations. Furthermore, regulations related to export controls can impose additional costs on the development and deployment of military AI systems. Again, these vary from country to country requiring in some cases a specialised team of experts on export controls to be organic to a company to assist in navigating multinational export contracts. Certain AI technologies may be subject to strict export controls, which can restrict the countries to which these technologies can be exported or shared. Navigating these regulations can require significant legal expertise and administrative resources. Breaches of these regulations can result in substantial penalties, including fines, sanctions, or even

prohibitions on the use of certain technologies. Beyond these direct costs, the changing nature of the regulatory landscape presents an ongoing challenge. As the implications and applications of AI continue to evolve, so too must the regulations govern its use. Staying abreast of these changes requires continuous monitoring and adaptation, further adding to the overall costs of regulatory compliance. Defence departments and AI developers must be prepared to adjust their policies, procedures, and systems in response to regulatory changes. This requires a level of agility and flexibility that can be costly to maintain but is crucial for ensuring long-term regulatory compliance. Overall, therefore, the cost component of regulatory compliance in the context of AI's integration into military systems can be significant. While the associated costs can be substantial, the implications of non-compliance, including potential fines, sanctions, and reputational damage, underscore the importance of investing in robust compliance mechanisms. The ability to navigate the complex and evolving regulatory landscape is not only a legal and ethical obligation but also a strategic necessity in leveraging the transformative potential of AI in the military domain responsibly and effectively.

Finally, it worth considering the risks introduced into critical supply chains through international venture capital flows and multinational business relationships. The multinational and dual-use of AI-enabled systems means that the ecosystem of commercial and research actors in the development of a given system are far broader than with more conventional modern military platforms [13]. This is particularly important in the case of autonomous systems and other military applications of AI because a failure point can be introduced (whether by accident or with malicious intent) at any stage of the process, for example in the collection, collation, and application of training data. As demonstrated by the finding of a Chinese-made alloy in the F-35 supply chain [39], the multinational web of companies involved in complex military industrial bases make it incredibly difficult for regulators to prevent supply chain intrusion. In this example it was simply an alloy, the security risk came from the potential for that firm to either input faulty parts or refuse supply. Contrastingly, the unknown participation of a compromised firm in the data training or base coding compilation for an AI technology could fundamentally undermine that system's reliability and effectiveness without necessarily leaving an identifiable trail. Focusing of excluding Chinese state-linked firms from sensitive supply chains (such as we have seen with Huawei [55]) risks overlooking another major potential source of advantage loss, either through acquisition, data leakage, or integration into adversaries' innovation networks leveraging access gained

through venture capital investments and corporate acquisitions [46]. As noted by Sayler (2020), there has been a significant "wave of investment" by US venture capitalists in AI that reached approximately $18.5 billion in 2019 [47]. Of note is that by 2015 Chinese venture capital investment was involved in 16% of all contracting in Silicon Valley firms [48], and by 2018 it had reached approximately 69% of the global total venture capital spending [49]. Bolstered by state-backed venture capital funds, the latter's investment in promising AI startups in places like Silicon Valley present not just IP challenges in the short term, they also lay the ground for longer-term sustainment security concerns.

4. Conclusion

In conclusion, the spike in public and policy maker interest in AI in mid-2023 represents an inflection point, an opportunity to adopt a systems approach to understanding the processes by which such innovations are translated into reality. As civilian actors consider the implications of democratised large learning models, militaries continue to pursue increasingly AI-enabled autonomous weapon systems. Both inventions represent potential demonstration points for disruptive, and potentially destructive uses of AI, and in both cases one must devote significant technological consideration and policy resources to understanding, mapping and addressing the often overlooked risks associated with developing and producing the underlying technology. The supply chains for such advanced systems are complex, multi-nodal, and cross-jurisdictional. Securing each stage from intrusion without artificially slowing innovation is a challenge particularly for democratic governments, which have more limited options for directing commercial actors. Such policy makers should be encouraged by the academic community to have meaningful discussions toward effective resilience building measures across the supply chain. Such resilience must be build early and reinforced in a multinational manner to ensure that future AI-enabled autonomous systems and other forms of military AI are developed, produced and deployed in a responsible and effective manner. Future avenues for research in this space would include evaluating mechanisms for collaborative development of AI safeguards for military systems, developing common concepts of operations for future deployment of autonomous systems in strategic logistics, and the potential for European Union members to develop integrated AI-enabled supply chains.

References

[1] A. Wyatt, *The Disruptive Impact of Lethal Autonomous Weapons Systems Diffusion: Modern Melians and the Dawn of Robotic Warriors*. Oxon and New York: Routledge, 2021.

[2] E. B. Kania, "Chinese military innovation in artificial intelligence," Testimony to the US-China Economic and Security Review Commission, 2019. [Online]. Available: https://www.cnas.org/publications/congressional-testimony/chinese-military-innovation-in-artificial-intelligence. [Accessed: Jan. 22, 2023].

[3] F. Sauer, "Stopping 'killer robots': Why now is the time to ban autonomous weapons systems," *Arms Control Today*, vol. 46, no. 8, pp. 8–13, 2016.

[4] Office of the Under Secretary of Defense for Policy, Directive 3000.09, United States Department of Defense, 2012.

[5] Development, Concepts and Doctrine Centre, Joint Concept Note 1/18: Human Machine Teaming, United Kingdom Ministry of Defence, 2018.

[6] Robotic and Autonomous Systems Implementation & Coordination Office, Robotic & Autonomous Systems Strategy v2.0, Canberra: Australian Army, 2022.

[7] M. C. Horowitz, "Why Words Matter: The Real World Consequences of Defining Autonomous Weapons Systmes," *Temple International and Comparative Law Journal*, vol. 30, pp. 85–98, 2016.

[8] P. Scharre, *Four Battlegrounds: Power in the Age of Artificial Intelligence*. New York: ww Norton, 2023.

[9] H. M. Roff, "The strategic robot problem: Lethal autonomous weapons in war," *Journal of Military Ethics*, vol. 13, no. 3, pp. 211–227, 2014, doi: 10.1080/15027570.2014.975010.

[10] I. Bode, H. Huelss, A. Nadibaidze, "Written Evidence AIW 0015," presented at the UK House of Lords AI in Weapon Systems Select Committee, 4 May 2023. [Online]. Available: https://committees.parliament.uk/writtenevidence/120184/pdf/. [Accessed: Jun. 6, 2023].

[11] L. Righetti, N. Sharkey, R. Arkin, D. Ansell, M. Sassoli, et al., "Autonomous weapon systems: technical, military, legal and humanitarian aspects," Proceedings of the International Committee of the Red Cross. Geneva, Switzerland, pp. 26–28, 2014.

[12] A. Wyatt, "Charting great power progress toward a lethal autonomous weapon system demonstration point," *Defence Studies*, vol. 20, no. 1, pp. 1–20, 2020, doi: 10.1080/14702436.2019.1698956.

[13] A. Ghadge, Maximilian Weiß, Nigel D. Caldwell, R. Wilding, "Managing cyber risk in supply chains: A review and research agenda," *Supply Chain Management: An International Journal*, vol. 25, no. 2, pp. 223–240, 2020, doi: 10.1108/SCM-10-2018-0357.

[14] S. Abaimov, M. Martellini, "Artificial intelligence in autonomous weapon systems," 21st Century Prometheus: Managing CBRN Safety and Security Affected by Cutting-Edge Technologies, pp. 141–177, 2020.

[15] M. C. Horowitz, *The diffusion of military power (The Diffusion of Military Power)*. Princeton, NJ: Princeton University Press, 2010.

[16] A. Wyatt, J. Galliott, "The revolution of autonomous systems and its implications for the arms trade," in *Research Handbook on the Arms Trade*, A.T.H. Tan, Ed. Cheltenham and Northampton, MA: Edward Elgar Publishing, 2020, pp. 389–405.

[17] A. B. Silverstein, "Revolutions in military affairs: A theory on first-mover advantage," B.A. thesis, University of Pennsylvania, Philadelphia, 2013.

[18] J. Kwik, "Mitigating the Risk of Autonomous-Weapon Misuse by Insurgent Groups," *Laws*, vol. 12, no. 1, 2023, doi: 10.3390/laws12010005.

[19] K. Chávez, O. Swed, "The proliferation of drones to violent nonstate actors," *Defence Studies*, vol. 21, no. 1, pp. 1–24, 2021, doi: 10.1080/14702436.2020.1848426.

[20] M. I. B. Amirruddin, "How Threat Assessments Can Become Self-Fulfilling Prophecies," *Pointer*, May, 2023.

[21] R. Gilpin, "The theory of hegemonic war," The Journal of Interdisciplinary History, vol. 18, no. 4, pp. 591–613, 1988, doi: 10.2307/204816.

[22] S. Shead, "UN talks to ban 'slaughterbots' collapsed – here's why that matters," in CNBC, ed, 2021.

[23] B. Zhang, M. Anderljung, L. Kahn, N. Dreksler, M. C. Horowitz, A. Dafoe, "Ethics and governance of artificial intelligence: Evidence from a survey of machine learning researchers," Journal of Artificial Intelligence Research, vol. 71, pp. 591–666, 2021, doi: 10.1613/jair.1.12895.

[24] A. Wyatt, J. Galliott, "An Empirical Examination of the Impact of Cross-Cultural Perspectives on Value Sensitive Design for Autonomous Systems," *Information*, vol. 12, no. 12, p. 527, 2021, doi: 10.3390/info12120527.

[25] J. Galliott, A. Wyatt, "A consideration of how emerging military leaders perceive themes in the autonomous weapon system discourse," *Defence Studies*, vol. 22, no. 2, pp. 253–276, 2022, doi: 10.1080/14702436.2021.2012653.

[26] A. Blanchard, M. Taddeo, "Autonomous weapon systems and jus Ad Bellum," AI & SOCIETY, pp. 1–7, 2022, doi: 10.1007/s00146-022-01425-y.

[27] E. Riesen, "The Moral Case for the Development and Use of Autonomous Weapon Systems," *Journal of Military Ethics*, vol. 21, no. 2, pp. 132–150, 2022, doi: 10.1080/15027570.2022.2124022.

[28] R. Waters, "Falling costs of AI may leave its power in hands of a small group," Financial Times, Mar. 10, 2023. [Online]. Available: https://www.ft.com/content/4fef2245-5559-4661-950d-6eb803fea329. [Accessed: Jun. 6, 2023].

[29] D. Nikolaiev, "Behind the Millions: Estimating the Scale of Large Language Models," 2023. [Online]. Available: https://towardsdatascience.com/behind-the-millions-estimating-the-scale-of-large-language-models-97bd7287fb6b. [Accessed: Jun. 6, 2023].

[30] M. DeGuerin, "'Thirsty' AI: Training ChatGPT Required Enough Water to Fill a Nuclear Reactor's Cooling Tower, Study Finds," in Gizmodo, 2023. [Online]. Available: https://gizmodo.com/chatgpt-ai-water-185000-gallons-training-nuclear-1850324249. [Accessed: Jun. 6, 2023].

[31] U. Gal, "ChatGPT is a data privacy nightmare. If you've ever posted online, you ought to be concerned," in University of Sydney News, 2023. [Online]. Available: https://www.sydney.edu.au/news-opinion/news/2023/02/08/chatgpt-is-a-data-privacy-nightmare.html. [Accessed: Jun. 6, 2023].

[32] B. Martin, L.H. Baldwin, P. Deluca, S. Henriquez, N. Hvizdaet et al., *Supply Chain Interdependence and Geopolitical Vulnerability: The Case of Taiwan and High-End Semiconductors.* Santa Monica: Rand Corp.

[33] K. Devitt, M. Gan, J. Scholz, R. Bolia, "A Method for Ethical AI in Defence," *Defence Science and Technology Group*, Contract No.: DSTG-TR-3786, 2021.

[34] T. Phillips-Levine, "The Art of Supply Chain Interdiction to Win Without Fighting," War on the Rocks, 2023. [Online]. Available: https://warontherocks.com/2023/05/the-art-of-supply-chain-interdiction-to-win-without-fighting/. [Accessed: Jun. 6, 2023].

[35] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, et al. "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," 2018. doi: 10.48550/arXiv.1802.07228.

[36] C. Strike, "Global Threat Report," *CrowdStrike*, 2023.

[37] V. Boulanin, *Mapping the development of autonomy in weapon systems: A primer on autonomy*. Stockholm: Stockholm International Peace Research Institute, 2016.

[38] A. Mehta, *Experiment over: Pentagon's tech hub gets a vote of confidence* [Online]. Available: https://www.defensenews.com/pentagon/2018/08/09/experiment-over-pentagons-tech-hub-gets-a-vote-of-confidence/. [Accessed: Jun. 6, 2023].

[39] L. Hudson, *Pentagon to resume F-35 deliveries after Chinese materials discovered*, Politico, 2022 [Online]. Available: https://www.politico.com/news/2022/10/07/pentagon-f-35-deliveries-chinese-materials-00060962. [Accessed: Jun. 6, 2023].

[40] R. Neuhard, *Foreign Policy Research Institute*, 2022. [Online]. Available: https://www.fpri.org/article/2022/10/the-new-us-national-security-strategy-four-take-aways-for-asia-policy/. [Accessed: Jun. 6, 2023].

[41] A. Holland Michel, "The black box, unlocked: predictability and understandability in military AI," United Nations Institute for Disarmament Research, 2020, doi: 1037559/SecTec/20/AI1.

[42] E. H. Christie, A. Ertan, L. Adomaitis, M. Klaus, "Regulating lethal autonomous weapon systems: exploring the challenges of explainability and traceability," AI Ethics, 2023, doi: 10.1007/s43681-023-00261-0.

[43] J. Haner, D. Garcia "The artificial intelligence arms race: Trends and world leaders in autonomous weapons development", *Global Policy*, vol. 10, no. 3, pp. 331–337, 2019, doi: 10.1111/1758-5899.12713.

[44] E. Schmidt, R. Work, S. Catz, E. Horovitz, S. Chien, A. Jassy, et al. "Final Report: National Security Commission on Artificial Intelligence (AI)," National Security Commission on Artificial Intelligence, Contract No.: AD1124333, 2021.

[45] A. Wyatt, J. Galliott, "Toward a Trusted Autonomous systems Offset Strategy: Examining the Options for Australia as a Middle Power," Australian Army Research Centre, Contract No.: 2, 2021.

[46] S. Korreck, "Exploring the Promises and Perils of Chinese Investments in Tech Startups: The Case of Germany," Observer Research Foundation, 2021.

[47] K. M. Sayler, "Artificial Intelligence and National Security," Congressional Research Service, Contract No.: R45178, 2020.

[48] M. Brown, P. Singh, "China's Technology Transfer Strategy: How Chinese Investments in Emerging Technology Enable A Strategic Competitor to Access the Crown Jewels of U.S. Innovation," Defense Innovation Unit – Experimental, 2018.

[49] E. H. Christie, C. Buts and C. Du Bois, "America, China, and the struggle for AI supremacy," 24th Annual International Conference on Economics and Security, Volos, Greece, July 8–9, 2021.

[50] M. C. Horowitz, "Artificial intelligence, international competition, and the balance of power," *Texas National Security Review*, vol. 22, 2018, doi: 10.15781/T2639KP49.

[51] M. Lamberth, P. Scharre, "Arms Control for Artificial Intelligence," *Texas National Security Review*, vol. 6, no. 2, pp. 95–110, 2023, doi: 10.26153/tsw/46142.

[52] S. Writer, "Fact Check-Simulation of AI drone killing its human operator was hypothetical, Air Force says," in *Reuters*, 2023. [Online]. Available: https://www.reuters.com/article/factcheck-ai-drone-kills-idusL1N38023R/ [Accessed: Dec. 4, 2023].

[53] E. Jones, B. Easterday, "Artificial Intelligence's Environmental Costs and Promise," in *Council on Foreign Relations*, 2022. [Online]. Available: https://www.cfr.org/blog/artificial-intelligences-environmental-costs-and-promise [Accessed: Dec. 4, 2023].

[54] L. Irwin, "How Much Does GDPR Compliance Cost in 2023?," in *IT Governance*, 2023. [Online]. Available: https://www.itgovernance.eu/blog/en/how-much-does-gdpr-compliance-cost-in-2020 [Accessed: Dec. 4, 2023].

[55] J.-Y. Lee, E. Han, and K. Zhu, "Decoupling from China: how US Asian allies responded to the Huawei ban," *Australian Journal of International Affairs*, vol. 76, no. 5, pp. 486–506, 2022, doi: 10.1080/10357718.2021.2016611.

[56] G. Baryannis, S. Validi, S. Dani, G. Antoniou, "Supply chain risk management and artificial intelligence: state of the art and future research directions," *International Journal of Production Research*, vol. 57, no. 7, pp. 2179–2202, 2019, doi: 10.1080/00207543.2018.1530476.

[57] R. Fedasiuk, J. Melot, B. Murphy, Harnessed lightning: *How the Chinese military is adopting artificial intelligence*. Washington DC: Center for Security and Emerging Technology, 2021.

[58] F. E. Morgan, M. Boudreaux, A. J. Lohn, M. Ashby, C. Curriden, et al., *Military applications of artificial intelligence*. Santa Monica: RAND Corporation, 2020.

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Guerre à la Carte: Cyber, Information, Cognitive Warfare and the Metaverse

Marco Marsili | Department of Philosophy and Cultural Heritage, Cà Foscari University of Venice, Italy; Centre for Research and Development, Military University Institute, Portugal, ORCID: 0000-0003-1848-9775

Abstract

Hybrid warfare is currently among the most trending topics. Hybrid threats arise in digital, cybernetic, and virtual environments and materialise in the real world. Despite being a somewhat vague term, hybrid activities include cyberwarfare, information warfare, and the emerging and evolving concept of cognitive warfare which appears from their intersection. These buzzwords gained popular attention in the context of the Russo-Ukrainian conflict and such terms are now in vogue. Even though these topics are in the spotlight, there is also widespread confusion about what exactly these usages mean and what the implications are in branding them as "warfare". Indeed, all these concepts are fluid, nebulous, and lack an undisputed legal definition. This article aims to clarify their meaning and to shed light on the characteristics of such terms – differences, similarities and overlaps – in the context of hybrid warfare and show the faulty reasoning upon which misunderstandings are based. The paper concludes with a glimpse into the future, closing with a reflection on multi-domain operations facilitated by a fully integrated human-computer interaction in the metaverse, where physical reality is merged and interacts with digital virtuality.

Keywords

cognitive, cyber, information, international humanitarian law, metaverse, warfare

Received: 24.04.2023

Accepted: 10.07.2023

Published: 13.07.2023

Cite this article as:
M. Marsili, "Guerre à la Carte: Cyber, Information, Cognitive Warfare and the Metaverse," ACIG, vol. 2, no. 1, 2023, DOI: 10.60097/ACIG/162861

Corresponding author:
Marco Marsili, Department of Philosophy and Cultural Heritage, Cà Foscari University of Venice, Italy; Centre for Research and Development, Military University Institute, Portugal; ORCID: 0000-0003-1848-9775; E-MAIL: marco_marsili@iscte-iul.pt

1. Introduction – The Nature of War

The nature of war has remained unchanged over time. Despite the popular quote attributed to Sun Tzŭ – "The nature of war is constant change" – the Chinese general never actually wrote this. On the contrary, in T*he Art of War, a tactical treatise for which he is traditionally credited as the author, Sun Tzŭ*, concludes that "in warfare, there are no constant conditions" [1, p. 53, § 32], which means, in the context of the text, that the battle is affected by ground, weather, and other contingent factors. In another overquoted classic book, *On War*, Clausewitz defines war as "an act of force to compel our enemy to do our will" [2, p. 75]. In his masterpiece, the Prussian general emphasises the use of "physical force" as an essential feature of war [2, p. 75]. As one of the most important treatises on political-military analysis and strategy ever written, even two centuries after its publication *On War* still influences strategic thinking. However, the core tenet of the book is undermined by misunderstandings and misleading interpretations [3, p. 90].

Kaldor [4, p. 221] argues that Clausewitz understood war as "the use of military means to defeat another state" and rejects this approach to warfare as no longer applicable in today's conflicts. She believes that current and future conflicts will not be ended through military victory, although violence remains a key feature. But the nature of war is always the same: defeat the enemy [1, p. 26–27]. A perfect summary of the nature of war is provided by Clausewitz himself: "[w]ar is more than a mere chameleon that slightly adapts its characteristics to the given case. As a total phenomenon, its dominant tendencies always make war a paradoxical trinity – composed of primordial violence, hatred, and enmity" [2, p. 89]. While the war on the battlefield is subject to specific conditions, which may change due to multiple factors, the nature of war is characterised by extreme violence and the use of weapons to overcome the enemy [2, p. 101, 3, p. 99, 5, p. 85, 6, pp. 68–69, 71–72].

Despite far too much rhetoric on the extension of the term "war" or "warfare", armed conflict is regulated by the legal framework provided by the Geneva Conventions, which define the perimeter of international humanitarian law (IHL), i.e., the "law of war" — IHL regulates the conditions for initiating war (*ius ad bellum*) and the conduct of waging parties (*ius in bello*), including occupation, and other critical terms of the law. Indeed, the wording "armed conflict" is relevant in the Conventions [7, p. 182, 8, p. 40–41]. Therefore, any use of the term "warfare" which does not involve the use of lethal weapons, is inappropriate [7, p. 191, 8, p. 45]. Due to overuse and misuse, "warfare" is now also applied to military operations other than war (MOOTW)

[9, p. 154, 8, p. 40–41]. Cyber-attacks may violate international law, when conducted or orchestrated by states, or may constitute cyber-crime, but certainly cannot be treated as kinetic attacks in the light of IHL [7, p. 191, 8, p. 42, p. 44–45]. Information warfare (IW) is not per se a change in the nature of war but rather a technological advance that can enhance lethal capacities [10, p. 16–19, 8, p. 44–45]. There is no evidence of any change to the nature of war [3, p. 91–92, 98, 10]. What changes is technology, along with techniques, tactics, and procedures [9, p. 152–156, 8, p. 37]. The topics of this paper should not be examined in isolation but should be seen as the first part of a larger argument. Nevertheless, there is an emerging doctrine that aims to characterise as "warfare" and/or "war" actions that are MOOTW; this trend mainly concerns "hybrid" operations, among which falls the cognitive domain.[1] That is why this premise is relevant to distinguish OTW activities from actions involving the use of actual force.

1 — See §3: Cognitive Warfare.

As M. L. R. Smith writes [11, p. 52], "Call it what you will – new war, ethnic war, guerrilla war, low-intensity war, terrorism, or the war on terrorism – in the end, there is only one meaningful category of war, and that is war itself" and Geneva Conventions apply. On these grounds one must reject the argument of Israeli military historian and theorist Martin van Creveld [12, p. 57–58] "[t]hat organized violence should only be called 'war' if it were waged by the state, for the state, and against the state". A state-centric approach to war is contradicted by the Conventions, which are crucial to this extent. Clausewitz conceptualised war as the application of violent means to realise military aims to achieve political ends, regardless of who the contenders are [3, p. 95].

2. Ruses of (Hybrid) Warfare

Foucault inverts Clausewitz's traditional conception of war and says that politics is the continuation of war by other means [13, p. 19]. Hybrid warfare is a concept that includes a wide range of tools – a bouquet of various techniques, methods, technologies, tactics, procedures and means, military and civilian, conventional and unconventional – for achieving a political or military objective [8, p. 37, 9, p. 151]. It is questionable whether *ruse de guerre* is legitimate or not. Misinformation, deception and electronic deception, electronic warfare, and psychological warfare are customarily accepted as lawful, and therefore they do not violate any general rule of international law applicable to armed conflict, so long as they do not involve treachery or perfidy [14, § 50–51, 15, §§ 8–3(b), 8–4(a), 8–5, 16, §§ 12.1, 12.1.1].

The European Union's definition of hybrid activities ranges from cyber-attacks through to disinformation; a combination of "coercive and subversive measures, using both conventional and unconventional tools and tactics (diplomatic, military, economic, and technological)" [8, p. 42–43]. The use of these tactics, aimed at targeting political institutions and influencing public opinion [9, p. 153, 155], is facilitated by rapid technological advances that reach a broad audience and which therefore boosts their impact.

NATO encompasses propaganda, deception, sabotage, and other non-military tactics among the hybrid methods of warfare [17]. The allies endorsed a vague definition of hybrid warfare at the 2016 Warsaw Summit: "a broad, complex, and adaptive combination of conventional and non-conventional means, and overt and covert military, paramilitary, and civilian measures" that are "employed in a highly integrated design by state and non-state actors to achieve their objective" [18, § 72]. The final communiqué issued at the 2021 meeting in Brussels groups cyber, hybrid, and other asymmetric threats, including disinformation campaigns, and sophisticated emerging and disruptive technologies [19, §§ 3, 12, 31].

While the Alliance has defined hybrid threats, the U.S. Department of Defense (DOD) has not officially provided a definition and has no plans to do so because hybrid warfare is not considered a new form of warfare since is a very broad term that blends conventional, unconventional, and irregular approaches across the full spectrum of conflict [20, p. 2, 11, 14].

Matuszczyk [21, p. 21] finds that these ruses of war, that go beyond conventional military capabilities, are simply creative, clever, unorthodox means. Bearing in mind that IHL sets the limits for acceptable wartime conduct (*ius in bello*), hybrid operations which do not involve the use of lethal force (despite being referred to as "warfare") fall below the threshold of armed conflict and cannot be characterised as such [8] — the lexicon and terminology are relevant to this end. If we accept that Clausewitz's famous statement that war is not merely an act of policy but a true political instrument, a continuation of political intercourse carried on with other means, we must therefore consider propaganda as a political tool [22, p. 23].

3. Cognitive Warfare

Although there is no common definition of hybrid warfare, the inclusion of propaganda, information and influence operations,

deception and psychological operations is widely accepted [9, p. 151]. Information warfare includes a set of techniques and technologies that ranges from electronic warfare to propaganda [9, p. 153], intertwined with the real and the virtual operational domains. The virtual realm encompasses electronic warfare (EW), electromagnetic spectrum operations (EMSO), cyberspace operations (CO), information warfare (IW), psychological operations (PSYOP), now better known as military information support operations (MISO), information operations (InfoOps or IO), also known as influence operations, strategic communications (STRATCOM), military deception (MILDEC), computer network operations (CNO), operations security (OPSEC), perception management (PM), public information (PI), and public diplomacy (PD) [9, p. 152–154].

Joint Publication 3–13, which provides doctrine and guiding principles for the U.S. Armed Forces, characterises IO as intended "to influence, deceive, disrupt, corrupt, or usurp the decision making" [23, § GL-3]. A 2018 U.S. Army pamphlet drafted by the Training and Doctrine Command (TRADOC) proposes the following definition of IW: "Employing information capabilities in a deliberate disinformation campaign supported by actions of the intelligence organizations designed to confuse the enemy and achieve strategic objectives at minimal cost" [24, § GL-6]. The publication highlights the relevance of information environment operations (IEO) and the convergence between the physical, virtual, and cognitive dimensions [24, §§ 3–3(d), 3–8(e), C-1].

The Information Environment (IE) impacts on the three dimensions (physical, virtual, cognitive). The fact that most cognitive activities occur primarily in the virtual domain does not mean that they have no effects in the real world. We can distinguish between two types of information disruption. The first is cognitive disruption, which includes any action (e.g., disinformation and propaganda) that directly targets individuals. The second is a functional disruption (e.g., cyberspace and electromagnetic attack) that directly targets systems and facilities (e.g., computers, weapons, vehicles) [25, § 3–15].

A U.S. Marine Corps publication introduces a conceptual framework on the ever-changing information environment in all warfighting domains, and highlights that information is "the foundation of all human interaction", accelerated and expanded by technologies "with a tempo and scale previously unimaginable" [25, Foreword]. The booklet quotes Sun Tzu's maxim "All warfare is based on deception" and acknowledges the relevance of deception defined as "an information activity [...] to deceive the human mind, the machine the human

relies on, or both" [25, §§ 2–22, 2–23]. The human-machine interaction is a fundamental component of cognitive warfare (CogWar) and plays a central and crucial role, due to the way our perception and judgment are affected, thus making it an unprecedented challenge.

Today's world is characterised by the widespread use of mobile digital communications and media which operate in largely ungoverned digital spaces [25, § 3–18]. The intersection of the information, physical and cognitive/social domains [9, p. 152], empowered by the digital ecosystem – the Internet, social media, and communication applications – creates the conditions for cognitive operations. Though there is nothing new among its individual components, the novelty in CogWar is the speed and power of dissemination of beliefs – false or true – instilled deeply in the consciousness of targets. The "infodemic" that arose in the context of the COVID-19 pandemic [26] can serve as a touchstone. This blurring effect makes people unknowingly susceptible to placing undue trust in specific information and sources or withholding it altogether due to outright confusion.

As human cognition is highly susceptible to manipulation and deception, CogWar aims to influence thinking processes, such as perceptions, decision making and behaviour [25, § 2–19]. Recognising and dispelling misinformation and disinformation requires critical thinking skills to identify untrustworthy information sources, and to understand how one's own potential cognitive biases may increase one's susceptibility to manipulation or influence [25, § 2–15]. This weaponised use of information serves to build and reinforce biased or false narratives, altering the perception and the behaviour of individuals and ultimately that of society [9, pp. 162–165]. Indeed, CogWar targets influential individuals, specific groups, and large numbers of citizens selectively and serially within society, with the potential to fracture and fragment an entire society or disrupt alliances [27].

In short, CogWar is a form of propaganda spread through manipulated media or social media for political or military purposes and aimed at fostering and instilling biased and conflicting narratives among targeted individuals, so as to make them behave accordingly by clouding their judgement. Therefore, what is most concerning about the cognitive effects of CogWar in peacetime is not its impact on the battlefield but the political and social consequences.

Cognitive science is the study of the human mind and brain, focusing on how the mind represents and manipulates knowledge and how mental representations and processes are actualised in the brain. Its

interdisciplinary features – linguistics, psychology, neuroscience, philosophy, computer science/artificial intelligence, anthropology, and biology – make cognitive science an autonomous academic discipline which studies the mind and its processes from different perspectives and approaches. It deals with human behaviour, with a focus on the mind and its interactions with the surrounding world, and how nervous systems represent, process, and transform information, and therefore is crucial to understanding the relevance and the impact of CogWar on brain, mind, and behaviour.

There are different views regarding the definition and intended scope of cognitive science, which can be considered "a multidisciplinary endeavour" that integrates methods and theories [28]. Paul Thagard [29] connects the origins of cognitive science to the first studies about the nature of human knowledge, of mind and mental operations, and to experimental psychology, linking them to the mid-50s, when primitive computers appeared, and artificial intelligence (AI) started to become conceptualised. In such context, Arthur Samuel [30] coined the term "machine learning" in 1959, following the publication (1950) of Alan Turing's seminal paper "Computing Machinery and Intelligence" [31]. Since then, AI, which comes from a deep learning approach based on neural networks, has become a central part of cognitive science [29].

While on the one hand it is clear that cognitive science is deeply interconnected to the human mind, on the other hand in order for it to be an autonomous discipline it needs an artificial – electronic and digital – environment provided by computers. Against this background, artificial intelligence and machine learning play a fundamental role, along with digital multimedia platforms, that empower global interconnectivity.

Thagards [29] finds that people have mental rules and procedures for generating new rules. CogWar techniques rely on such mental patterns and thereby influence the decision-making process and the behaviour of target populations by predicting and manipulating the results of perceptions and actions.[2] From these definitions and concepts we can infer the relevance and impact of CogWar, and the attention and concerns it raises.

The importance of CogWar and related topics is highlighted, *inter alia*, by the recent release (Sept. 2022) of the U.S. *Joint Publication 3.04 – Information in Joint Operations*, which provides fundamental principles and guidance to plan, coordinate, execute and assess the use of information during joint operations [32]. The revised doctrine,

2 —— For a discussion on behaviourism, see, e.g., G. Graham, „Behaviorism", in *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition), E.N. Zalta, U. Nodelman, Eds. [Online]. Available: https://plato.stanford.edu/archives/spr2023/entries/behaviorism/.

which has not been publicly released, briefly introduces cognition and its cognitive impact within the IE [32, §§ I-7, III-3, VI-2].

Both the U.S. joint doctrine and NATO policy have already recognised cyberspace as an operational military domain and are striving to include the cognitive realm among the battlefields [7, pp. 178, 181]. As the cognitive dimension becomes ever more relevant in the present and future geopolitical challenges, NATO is taking the necessary action against "weaponised information" in modern warfare. The NATO Allied Command Transformation (ACT) acknowledges that "the lines between peace and war, political and military, strategic and tactical, physical and non-physical are blurring" [33] and the Supreme Allied Command Transformation (SACT) Concept Development Branch (CNDV) has been accordingly tasked by SACT to develop a concept on cognitive warfare [27]. The work is part of the implementation of the NATO Warfighting Capstone Concept (NWCC) through the Warfare Development Agenda (WDA). The CogWar Concept is a Line of Delivery (LoD) nested under the cross-domain command of the Warfighting Development Imperative (WDI) [34, § 1], as identified by the NWCC.

A cognitive warfare exploratory concept is currently under development by a NATO ACT team of experts. The goal is to develop an Exploratory CogWar Concept for approval by SACT during 2023 in order to implement the NWCC and leverage the WDA. This exploratory concept will include a final Cognitive Warfare Concept, to be approved by the Military Committee (MC) in the summer of 2024 – the MC develops strategic policy and concepts and provides guidance to SACT and as such is an essential link between the political decision-making process and the military structure of NATO, being tasked for translating political decision and guidance into military direction [35].

NATO's Military Strategy, adopted in May 2019, provides the Alliance with military-strategic objectives and the ways and means to implement them through two high-level concepts: the NWCC, as part of the WDA – a planning tool to implement the Warfighting Capstone Concept – and the Allied Command Operations (ACO) Concept for the Deterrence and Defence of the Euro-Atlantic Area (DDA) [33; 34]. Endorsed by NATO Heads of State and governments in 2021, the NWCC, often referred to as NATO's North Star, sets forth a 20-year vision by anticipating threats and understanding the strategic environment and specifically focuses on multi-domain operations (MDO), resilience, cognitive work and much more, enabled by digital transformation [33; 34]. MDO are how operations are conducted in time and space with synchronisation of all domains [36] and are described by TRADOC

as a mix of "unconventional and information warfare (social media, false narratives, cyber-attacks)" [24, vi, §§ 2–2, C-2, D-3].

According to the definition developed by the NATO team of experts, "'Cognitive Warfare' is the convergence of 'Cyber-Psychology', 'Weaponization of Neurosciences', and 'Cyber-Influence' for a provoked alteration of the perception of the world and its rational analysis by the military, politicians, and other actors and decision-makers, to alter their decision or action, for obtaining strategic superiority at all levels of tactical intervention concerning individual or collective natural intelligence, as well as artificial or augmented intelligence in hybrid systems" [8, p. 44].

The NATO Science and Technology Organization (STO) has endorsed a variety of Exploratory Teams (ET) and Research Task Groups (RTGS) on the subject of CogWar [37].[3] The System Analysis and Studies (SAS) Panel approved the following RTGS: SAS-177 on Defending Democracy in the Information Environment: Foundations and Roles for Defence; SAS-185 on Indicators and Warnings for Cognitive Warfare in Cyberspace. The Information Systems Technology (IST) Panel endorsed the following activities: IST-177 (RTG) on Social Media Exploitation for Operations in the Information Environment; IST-ET-123 on Exploring Countermeasures against Misinformation of a Nation's Population. Interdisciplinary research led by the Human Factors and Medicine (HFM) Panel include: HFM-374-RTG CogArmy: Cognitive training and teamwork assessment of Army personnel; HFM-ET-214 Cognitive Security: building and maintaining resistance to offensive cognitive strategies; HFM-ET-215 The Ethical and Legal Challenges of Cognitive Warfare; HFM-ET-216 Methods and Weapons of Adversary Cognitive Warfare; HFM-IST-ET-213 Visualization of Cognitive Warfare Situational Awareness; HFM-373-RTG Technology Enablers for Monitoring and Assessment of Humans in CogWar. These research activities were approved by different panels — which reinforces the cross-disciplinary of the topic (a good example of this is the SAS-HFM-ET-FE on Early Warning System for Cognitive Warfare in Cyberspace). Most of this research activity is classified or restricted and not publicly releasable and therefore we will not dwell on such content in this article.

In this context, the NATO STO Human Factors & Medicine Panel organised an HFM-361 Research Symposium (RSY) on Mitigating and responding to Cognitive Warfare in Madrid on 13–14 November 2023, aimed at supporting the WDA (as stated in the NATO NWCC) and providing information for science and technology guidance on improving countermeasures to CogWar, so as to meet and mitigate current

2 — Situation updated as of 5 July 2023.

and future security and defence challenges [38]. The proposal for a symposium on Meaningful Human Control in Information Warfare (HFM-377-RSY) to be held in the coming year is still pending.

4. The Metaverse: A New Domain of Warfare?

The term "metaverse" – a portmanteau which combines the words "meta" (meaning beyond) and "verse" (short for the universe) – increases the confusion on and around defence concepts that lack a workable definition. This hip buzzword was coined in 1992 by visionary author Neal Stephenson in his dystopian sci-fi thriller *Snow Crash* [39], which predicted the metaverse as a convergence between the real and the virtual world; a universe beyond the physical where physical reality is merged and interacts with digital virtuality [40, p. 486] facilitated by the Internet of Things (IoT). According to one of the many similar definitions, IoT "is the network of physical objects that contain embedded technology to communicate and sense or interact with their internal states or the external environment" [41].

The two words – metaverse and war – may sound completely unrelated but on closer consideration they are more intertwined than they may appear at first glance. The virtual and physical worlds are becoming increasingly interconnected, interdependent, and indistinguishable from one another. Metaverse wars draw together online and offline worlds. In traditional warfare, enemies go to war with (or over) something tangible. Since cyberspace was elevated to the domain of operations, just as for the three traditional realms of land, air, and sea [7, pp. 178, 181], cyberspace became a virtual battlefield. This new way of waging war where opponents can do battle in a virtual environment could replace physical wars.

What happens in cyberspace does not necessarily stay in cyberspace. The metaverse can serve as a bridge to bring the actual force from the virtual to the real world, going far beyond the boundaries of a traditional conflict. As Stephenson wrote, "The Metaverse has now become a place where you can get killed" [39, p. 346] — a fictional statement which genuinely raises concerns. Kinetic actions can be materialised through cyberspace and reverberate their effects in the classic operational and physical domains. However, until cyber actions involve the use of lethal force, they fall below the threshold of armed conflict [7, p. 189–191; 8, p. 40–41].

Even if virtual actions cannot replace physical warfare as such, it does not mean that cyberwarfare has no negative impact. A drone attack

conducted virtually can have lethal effects on the battleground. As war becomes the counterpart of communication, the latter unfolds its effects, even if not lethal, in the real world. This implies that a nation's power would no longer be decided just by its resources and manpower, but by its critical enabling capabilities across all domains. Stephenson writes that everything in the metaverse "depends upon the ability of different computers to swap information very precisely, at high speed, and at just the right times" and that "people who go into the Metaverse…understand that information is power" [39, p. 400, 431]. If we connect these fictional words to the real world, we can easily imagine the impact of the metaverse on military operations, where the convergence between cyberoperations and electromagnetic operations plays a crucial role in gaining full spectrum dominance [42].

The concept of "full spectrum operations" highlights the influence of full-dimension operations on future doctrine [42]. Given the cross-domain, multi-domain, or all-domain operations doctrine, which prompts the military to conduct full spectrum operations to exert control over all dimensions of the battlespace [24], it seems clear that the metaverse may result in a new domain of warfare over time, although it is still too early to say how. What is also clear is the legal framework, which should be respected.

The significance of the interconnection between the cyber domain and the metaverse for multi-domain operations is confirmed by research commissioned by the Italian Ministry of Defence, in the scope of the Annual Research Plan (2023), with the purpose of identifying and exploring dual-use and innovative technologies to enhance military capabilities and gain a tactical advantage, in line with NATO STO trends [43].

While digital transformation enables MDO, emerging and disruptive technologies – including, *inter alia*, virtual and augmented reality – have further complicated the operational environment. The multi-domain environment can be dubbed the "metaverse", an immersive visual interaction between physical and virtual objects facilitated by advancing virtual reality (VR) and haptic technology [44, p. 97, 99]. The metaverse is bringing the physical and digital worlds closer together by expanding the possibilities of virtual and mixed reality and finally interacting with the physical and digital worlds. Potential applications in the metaverse include building and manipulating 3D objects and creating more intuitive, human-centred interfaces through AI. [44, p. 94].

The next generation of wearable technologies – textile computing technologies that can sense and react to the human body – will

enhance the experience of the users to provide a fully integrated human-computer interaction through digitisation of human biodata, activities, behaviours, and relationships, turning textiles into bidirectional interfaces that might find effective military applications [44, p. 99].

5. Conclusions

Emerging and evolving threats are coming from the virtual and cyber domains. Even if this appears to be nothing new, what is novel is the speed, scale and intensity of unconventional attacks, facilitated by rapid technological change and global interconnectivity. It is more than likely that such threats will increase in the future until they become prevalent over conventional (kinetic) means of warfare, although rapid technological advance and emerging military doctrine prevent us from reaching any definite conclusion at this point. Future research should scrutinise the impact of cognitive actions and the metaverse on individuals – a broad audience encompassing political and military leaders, policy and decision-makers and the society as a whole – and how international relations and warfare may be affected.

While rapid technological change makes the future of warfare uncertain and unpredictable, the metaverse seems to have the potential to become a new battlefield where information and cognitive operations could find their "natural" environment. Nevertheless, such operations are lawful either in the real or the virtual world; the emerging military doctrine cannot equate non-kinetic and non-lethal actions to a conventional attack.

"If we hold to the assertions by Sun Tzŭ, Clausewitz, Smith and Foucault, we must conclude that, while there is no distinction between political and military activity, the latter is characterised by the use of lethal weapons, and any other activity has to be considered as below the threshold of armed conflict and outside the scope of war(-fare) according to IHL, including information and cognitive actions and, to some extent, cyberattacks, with the metaverse that, given its hybrid nature, can support either kinetic and non-kinetic operations".

Although the legal framework is clear, governments and military organisations should strive to reach a legally binding and undisputed definition of threats coming from the digital world, whilst taking care not to brand them as "warfare" so as to avoid triggering any conventional response. International law cannot be made through one party's doctrine or policy. Peace is the most valuable commodity and is too precious to be endangered by virtual conflicts.

Funding and Acknowledgement

The author gratefully acknowledges the Ministry of University and Research (MUR), Italy, for supporting his work through the Young Researchers-Seal of Excellence (SOE) grant funded by NextGenerationEU (NGEU) under the National Recovery and Resilience Plan (NRRP).

References

[1] Sun Tzŭ, *The Art of War. London:* Luzac & Co., 1910.

[2] C. von Clausewitz, *On War.* Princeton: Princeton University Press, 1984.

[3] B. Schuurman, "Clausewitz and the 'New Wars' Scholars," The US Army War College Quarterly: Parameters, vol. 40, no. 1, pp. 89–100, 2010, doi: 10.55540/0031-1723.2515.

[4] M. Kaldor, "Elaborating the 'New War' Thesis," in *Re-thinking the Nature of War*, I. Duyvesteyn, J. Angstrom, Eds. New York: Frank Cass, 2005, pp. 210–224.

[5] L. Freedman, "War Evolves into the Fourth Generation: A Comment on Thomas X. Hammes," in *Global Insurgency and the Future of Armed Conflict: Debating Fourth-Generation Warfare*, T. Terriff, A. Karp, R. Karp, Eds. New York: Routledge, 2008, pp. 78–86.

[6] M. Evans, "Elegant Irrelevance Revisited: A Critique of Fourth Generation Warfare," in *Global Insurgency and the Future of Armed Conflict: Debating Fourth-Generation Warfare*, T. Terriff, A. Karp, R. Karp, Eds. New York: Routledge, 2008, pp. 67–74.

[7] M. Marsili, "The War on Cyberterrorism," *Democracy and Security*, vol. 15, no. 2, pp. 172–199, 2019, doi: 10.1080/17419166.2018.

[8] M. M. Marsili, "Hybrid Warfare: Above or Below the Threshold of Armed Conflict?," *HSZ-HDR*, vol. 150, no. 1–2, pp. 36–48, 2023, doi: 10.5281/zenodo.7557494.

[9] M. Marsili, "The Russian Influence Strategy in Its Contested Neighbourhood," in *The Russian Federation in Global Information Warfare. Influence Operations in Europe and Its Neighborhood*, H. Mölder, V. Sazonov, A. Chochia, T. Kerikmäe, Eds. Cham: Springer, 2021, pp. 149–172, doi: 10.1007/978-3-030-73955-3_8.

[10] C.S. Gray, "The Changing Nature of Warfare?," *Naval War College Review*, vol. 49, no. 2, pp. 7–22, 1996.

[11] M.L.R. Smith, "Strategy in the age of 'low-intensity' warfare: why Clausewitz is still more relevant than his critics," in *Re-thinking the Nature of War*, I. Duyvesteyn, J. Angstrom, Eds. New York: Frank Cass, 2005, pp. 28–64.

[12] [12] M. van Creveld, On Future War. London: Brassey's, 1991.

[13] M. Marsili, "From Battlefield to Political Arena. Shifting the Clausewitzian Paradigm," *Political Reflection*, vol. 7, no. 2 (issue 27), pp. 19–25, 2021, doi: 10.5281/zenodo.4554695.

[14] U.S. Department of the Army, Field Manual 27–10, The Law of Land Warfare. Washington, DC: U.S. Department of the Army, 18 July 1956, as modified by Change No. 1, 15 July 1976.

[15] U.S. Department of the Air Force, Air Force Pamphlet 110–31, International Law – The Conduct of Armed Conflict and Air Operations. Washington, DC: U.S. Department of the Air Force, 1976.

[16] U.S. Navy, U.S. Marine Corps, U.S. Coast Guard, The Commander's Handbook on the Law of Naval Operations, NWP 1-14M/MCTP 11-10B/ COMDTPUB P5800.7A. Department of the Navy, Office of the Chief of Naval Operations and Headquarters, US Marine Corps, and Department of Homeland Security, U.S. Coast Guard, Edition March 2022.

[17] NATO, "NATO's response to hybrid threats," Apr. 4, 2023. [Online]. Available: https://www.nato.int/cps/en/natohq/topics_156338.htm. [Accessed: Apr. 7, 2023].

[18] Heads of State and Government. (2016, July 9). Warsaw Summit Communiqué. [Online]. Available: https:// www.nato.int/cps/en/natohq/official_texts_133169.htm. [Accessed: Apr. 18, 2023].

[19] Heads of State and Government. (2021, June 14). Brussels Summit Communiqué. [Online]. Available: https://www.nato.int/cps/en/natohq/news_185000.htm. [Accessed: Apr. 18, 2023].

[20] U.S. Government Accountability Office. (2010, Sep. 10). Hybrid Warfare, GAO-10-1036R. [Online]. Available: https://www.gao.gov/products/gao-10-1036r. [Accessed: Apr. 17, 2023].

[21] A. Matuszczyk, *Creative Stratagems: Creative and Systems Thinking in Handling Social Conflict*. Kibworth: Modern Society Publishing, 2012.

[22] M.M. Marsili, "Propaganda and International Relations: An Outlook in Wartime," *ArtCiencia.com*, no. 19, pp. 1–26, 2015, doi: 10.25770/artc.11095.

[23] Joint Chiefs of Staff (JCS). JP 3–13, Information Operations (Joint Publication 3–13), Incorporating Change 1, 20 November 2014. Washington, DC: JCS, 27 November 2012.

[24] U.S. Training and Doctrine Command (TRADOC). (2018, Dec. 6). The U.S. Army in Multi-Domain Operations 2028. [Online]. Available: https://adminpubs.tradoc.army.mil/pamphlets/TP525-3-1.pdf. [Accessed: Apr. 20, 2023].

[25] U.S. Marine Corps, Information, MCDP 8. Washington, DC: Department of the Navy, 2022.

[26] M. Marsili, "COVID-19 Infodemic: Fake News, Real Censorship. Information and Freedom of Expression in Time of Coronavirus," *Europea*, vol. 10, no. 2, pp. 147–170, 2020, doi: 10.4399/97888255402468.

[27] K. Cao, S. Glaister, A. Pena, D. Rhee, W. Rong, A. Rovalino, S. Bishop, R. Khanna, and J. Singh Saini, "Countering cognitive warfare: awareness and resilience," NATO Review, May 20, 2021. [Online]. Available: https://www.nato.int/docu/review/articles/2021/05/20/countering-cognitive-warfare-awareness-and-resilience/index.html. [Accessed: Apr. 8, 2023].

[28] R. Núñez, M. Allen, R. Gao et al., "What happened to cognitive science?," *Nat Hum Behav*, vol. 3, pp. 782–791, 2019, doi: 10.1038/s41562-019-0626-2.

[29] P. Thagard, "Cognitive Science," 1996. [Online]. Available: https://plato.stanford.edu/archives/spr2023/ entries/cognitive-science/. [Accessed: July 4, 2023].

[30] A.L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," IBM J. *Res*. Dev., vol. 3, no. 3, 1959, pp. 210–229, doi: 10.1147/rd.33.0210.

[31] A.M. Turing, "Computing Machinery and Intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1959, doi: 10.1093/mind/LIX.236.433.

[32] Joint Chiefs of Staff (JCS). JP 3–04, Information in Joint Operations (Joint Publication 3–04). Washington, DC: JCS, 2022.

[33] NATO ACT, "The Alliance's Warfare Development Agenda: Achieving a 20-year Transformation," Mar. 29, 2022. [Online]. Available: https://www.act.nato.int/articles/wda-achieving-20-year-transformation. [Accessed: Apr. 16, 2023].

[34] NATO ACT, "NATO Warfighting Capstone Concept (NWCC)," [Online]. Available: https://www.act.nato.int/nwcc. [Accessed: Apr. 16, 2023].

[35] NATO, "Military Committee," Oct. 3, 2022. [Online]. Available: https://www.nato.int/cps/en/natohq/topics_49633.htm. [Accessed: Apr. 20, 2023].

[36] M. Marsili, "Shaping the Holistic Concept of Multi-Domain in a Legal Vacuum. A Tricky Issue". in *II Encontro Anual da Investigação & Desenvolvimento em Ciências Militares (EAI&DCM2019)*, 2019. doi: 10.5281/zenodo.3473959.

[37] HQ Supreme Allied Commander Transformation Strategic Plans and Policy Directorate. (2022, Sep. 29). Cognitive Warfare (CW) Initial Validation Planning Workshop 27–28 October 2022, ACT/SPP/CNDV/TT-4563/ SER:NU.

[38] NATO STO. HFM-361 on Mitigating and responding to cognitive warfare. [Online]. Available: https://events. sto.nato.int/index.php/event-summary/event/17-symposium/507-hfm-361-on-mitigating-and-respondingto-cognitive-warfare. [Accessed: July 4, 2023].

[39] N. Stephenson, *Snow Crash*. New York: Bantam Books, 1992.

[40] S. Mystakidis, "Metaverse," Encyclopedia, vol. 2 no. 1, pp. 486–497, 2022, doi: 10.3390/ encyclopedia2010031.

[41] Gartner. Internet Of Things (iot). [Online]. Available: https://www.gartner.com/en/information-technology/ glossary/internet-of-things. [Accessed: July 4, 2023].

[42] Joint Chiefs of Staff (JCS), JP 3–0, Joint Operations (Joint Publication 3–0). Washington, DC: JCS, 2022.

[43] Istituto di ricerca e analisi della difesa – IRAD, "Avviso Esito Pubblico Ricerche 2023," May 24, 2023. [Online]. Available: https://irad.difesa.it/Comunicazione/Dettaglio/17. [Accessed: July 4, 2023].

[44] M. Marsili, "Epidermal Systems and Virtual Reality: Emerging Disruptive Technology for Military Applications," *Key Eng. Mater.*, vol. 839, pp. 93–101, 2021, doi: 10.4028/www.scientific.net/KEM.893.93.

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Cyberwarfare against Critical Infrastructures: Russia and Iran in the Gray Zone

Guillermo López-Rodríguez | Department of Political Science and Public Administration, University of Granada, Spain, ORCID: 0000–0001–8704–9007
Irais Moreno-López | Center of Political Studies, National Autonomous University, Mexico
José Carlos Hernández-Gutiérrez | Department of Political Science and Public Administration, University of Granada, ORCID: 0000–0002–2855–1053

Abstract

The holistic nature of security in a hyper-connected world has increased the relevance of the cyber environment. One of the most relevant threats identified are the attacks against energy infrastructure. This article establishes a comparative study of cyber--attacks launched by Russia and Iran against energy-related infrastructure. Both countries are specialized in asymmetric strategies and tactics in which cyber has a core role. The research analyses both Iran and Russia´s main actions against energy supply infrastructure, studying the pursued objectives and identifying their potential political results. The document is structured as an initial theoretical approach to the use of asymmetric Gray Zone and hybrid strategies, focusing on the use of cyber-attacks by Rogue States. From this approach, the analysis reflects the political visions of Russia and Iran, linking it with Russia's actions in Ukraine, as well as the Iranian cyber offensives against western targets. The concluding section reflects on the effectiveness of these strategies with respect to the general strategy of both states.

Keywords
Iran, Russia, Cyberwarfare, Hybrid, Gray Zone

Received: 29.06.2023

Accepted: 16.10.2023

Published: 31.12.2023

Cite this article as:
G. López-Rodríguez, I. Moreno-López, J. C. Hernández-Gutiérrez "Cyberwarfare against Critical Infrastructures: Russia and Iran in the Gray Zone," ACIG, vol. 2, no. 1, 2023, DOI: 10.60097/ACIG/162865.

Corresponding author:
Guillermo López-Rodríguez, Department of Political Science and Public Administration, University of Granada, Spain; ORCID: 0000–0001–8704–9007; E-MAIL: guillermolopez@ugr.es

1. Introduction

The geopolitics of energy in the 20th and 21st Centuries was controlled by the power of the Oil-States to be able to shut down the supply of oil and gas. The progressive energy transition has reduced the dependency on fossil fuels, and renewable energies have less scope to be used a tool of geopolitics as energy production becomes more decentralized. However, new threats are emerging, one of them being the supply interruptions due to cyber-attacks against critical infrastructure. One relevant example is the Russian attack against Ukraine in 2015 that left 250,000 people without energy supply [1].

Among the critical infrastructures the electricity network is one of the most relevant, due to the dependence of telecommunications, transport, the financial system or public security upon them [2]. Without electricity, the financial sector, emergency services and public institutions could be disrupted [3]. A disruption to the electrical network could also have fatal consequences: as an immediate result of the shutdown there could be dead and wounded, owing to fire, hypothermia, gas leaks, failures in the healthcare system or interruptions to the water supply [4]. The high level of interconnection between technological networks may pose a threat to state cybersecurity, as has been demonstrated in recent years with the cyber-attacks against national networks in certain countries [5].

Electricity networks are a priority target of the military and insurgents. For terrorist groups and organised insurgencies, it is cheap and easy to destroy high-voltage pylons or attack power stations. On the other hand, it is typical in military strategy is usual to plan kinetic offensives against power plants or analogous installations as part of bombing campaigns. Cyber-attacks are part of the portfolio of strategies of states [6]. It can be stated that the energy sector is exposed to a wide variety of attacks, with some of them falling within the framework of hybrid warfare [7]. In this sense, the United States has identified Russia, China, Iran and North Korea as critical threats to its energy sector [2].

This article analyses the cyber actions launched by Russia and Iran against the energy supply infrastructure of their adversaries. Both states are specialised in asymmetric strategies and tactics, defined by their lack of recognition, ambiguous objectives, and the use of proxy non-state actors. This is a comparative study that focuses on actions developed in order to achieve their geopolitical imperatives.

Due to this fact, the main objectives in this paper are to analyse the cyber actions as part of a broader strategy, with the further aim of

studying the most relevant cyber actions conducted against energy supply infrastructure. One of the most important contributions of the paper is to establish a link between the concept of the 'Gray Zone' and hybrid warfare, and applying this to the specific cases of Russia and Iran.

Starting from our approach to the use of asymmetric warfare by state actors, the analysis contains an outline of the geopolitical visions of Russia and Iran, so as to facilitate examination of their actions within the cyber domain. This research focuses on the Russian attacks against Ukraine and the Iranian cyber operations against Western targets. The conclusion reflects upon the effectiveness of these methods with reference to the general strategies of both states.

2. The Gray Zone and cyber strategies

Recent decades have shown the need for states to augment their military capabilities with more subtle ways of exerting their power [8]. For instance, power can be exercised incentives, bribery or coercion [9]. The liberal international trading environment provides opportunities for rapid economic development, while also providing illiberal states with ways of exploiting this environment in their favour to increase their relative power by evaluating their network sources [10]. While conventional warfare utilises a high number of communicative components, non-conventional actions are more difficult to capture and identify, combining economic, cultural or technological features in a geopolitical project [11].

In this respect, the use of energy as a geopolitical tool has become a key asset for shaping international relations [12]. This use might be driven by economic motives, political reasons, or even national security matters [13]. Using energy as a political and military resource has allowed states to influence other countries and their decisions by controlling energy supply or demand [14], as well as influencing control and access to their own resources or supply [15]. The availability of resources is considered a key element that determines the geopolitical behaviour of states [16], such that it can be turned into a target to be attacked conventionally and unconventionally by their adversaries.

Due to its relevance and the potential damage caused by its shortage, energy is instrumentalized in Gray Zone conflicts. The spectrum of competition space existing between states is positioned between the

polarities of absolute peace and conventional armed conflict [17]. The Gray Zone is a type of conflict in which actors seek limited political victories, acting within a murky environment which does not explicitly break the rules and values of the international world order [18]. The strategy implies a gradual conflict that seeks to modify components of the international system, using a combination of soft and hard power measures in non-conventional ways, making it difficult to prevent or respond to them [19]. This is a long-term approach, coordinating state and non-state actors and seeking to generate deterrence regarding the adversary [17].

The Gray Zone implies a temporary sustained confrontation that would not escalate to full-scale conventional war. The tools employed are economic and political pressure, energy blackmail or the use of cyberwarfare, both to reduce escalation of violence and to prevent retaliation [19]. This process is defined by a lack of clarity for the adversary, using ambiguity to weaken deterrence. However, both the revisionism of the international order and its alliances are a relevant component of the strategic objectives implied in this way of conflict [10]. The permissive and advantageous conditions are created by illiberal states such as China or Russia to effectively conduct operations in the Gray Zone against democratic countries. The lack of legal regulations allows authoritarian states to normalise new practices and tools in this Gray Zone. Authoritarianism is highly centralised yet bureaucratically flexible, which allows for more effective use of propaganda, legal national structures, economic pressure, and support for non-state proxies, in comparison to what democracies can employ.

The success of Gray Zone operations is based in the interconnection of political, informational, and economic domains [10]. In a more digitalised world, the cyber dimension has increased its relevance [17]. Cyber operations contribute to increasing the opacity of actions, due to the difficulties in assigning responsibility [20]. Due to this fact, cyberwarfare is closer to terrorism and guerrilla warfare than conventional warfare, being considered in some cases a force multiplier or a strategic tool in others [11]. The effects are even more pronounced, considering their low cost, disruptive potential and the high level of damage can be inflicted upon an adversary [20].

According to the literature, those states with aggressive geopolitical agendas are likely to enhance their virtual and automated tools in the future [21]. The cyber actions enable and require cooperation between public and private actors [17], which hinders the attribution of responsibility, while diversifying the objectives and increasing the

potential effects of network operations [22]. Cyberwarfare increases the attacker's advantages due to the element of surprise, which can hamper the opponent's ability to react when combined with other types of actions [11]. One of the key objectives can be critical non-military infrastructures, with unpredictable domino effects leading to scenarios that can lead to long-term blackouts with impacts on other services, such as healthcare or food supply [22].

3. The Russian and Iranian world view

Historically, Russia and Iran had maintained a strategic relationship towards a common enemy which they both conceptualize broadly as "the West"; this means, the United States and the European Union. The conceptual differentiation between Russia and The West has been a core element in the Russian cultural and philosophical tradition since the end of the 19th Century (Berlin, 1978). In the case of Iran, the 1979 Islamic Revolution formed the milestone for taking a religious, ideological and political distance from the West. The gulf expressed by both countries' leadership implies a different approach to the reality having its translation into politics [23]. This strategic agreement does not involve a dovetailing of ideological or moral perspectives between the leaderships of these two countries, but a common defensive view from their foreign policy standpoint and national interests that emphasizes the need to fight what they regard as western impositions, both in international, political and economic scenarios as well as in social life. There are two important elements that help to explain how the view of foreign policy in Iran and Russia is related to actions that lead to these kind of cyber-attacks on critical infrastructure, particularly on energy.

First, a traditional defensive view of Iran and Russia from what they call "the West" as a *de facto* international power headed formally and informally by the US and EU. This perception led to a specific view of their role in the international scenario, where "the West" is constantly trying to impose upon the rest of the world those views, lifestyle and policies that serve western interests. The transformation of international relations from a bipolar system into multipolarity implied that Russian elites would seek to be one of these poles [24]. Second, the view of both countries of foreign policy as a zero-sum game. Russia developed a particular combination of these two elements. Since Vladimir Putin's second term (2004–2008), Russia has re-emerged with the same policy stance as that adopted during the Cold War: once the menace from terrorism stopped as a common cause for United States and Russia, Vladimir Putin started

to distance himself from the West and began to claim the historical role of Russia – as heir of the USSR- in international politics. From that perspective, the very secure and already popular regime of the Russian president began to speak about a global shared leadership between the great powers, as existed from 1945–1991. The Russian invasion of Georgia in 2008 implied a rupture in the geopolitical relations between Western and Eastern countries. Prior to this military offensive, political analysts had assumed that the economic transformation in Russia would imply a closer approach to the West. In addition to the Georgian invasion, other key elements that widened the East-West gap were the Iranian nuclear program and the Afghan War [25].

The doctrine that nourished this perspective was called "Eurasianism", a long-forgotten term for this part of the world that no longer felt like part of Europe, and was not exclusively part of Asia either. This doctrine of Eurasianism forcefully vindicates the role of Russia as a great international power, that is, a kind of rationale for no western "intervention" in what they call the near abroad countries (meaning former USSR republics). The doctrine claims Russia's regional leadership of a symbolically constructed region called Eurasia in the name of power sharing across different regions of the world. According to Marcin Skladanowski [26], it was Dugin who founded "the Myth of Russian Exceptionalism", which is described as follows:

> *"The conviction of Russia's uniqueness, both in the past as well as the present" and this uniqueness has become the fuel to radicalize the anti-Western rhetoric of the Russian Federation because of its anti-Occidental identity awakening* [26].

In a similar way, from the moment the religious movement of Ayatollah Khomeini succeeded in Iran in 1979, a campaign of radical anti-westernization was undertaken by the new theocratic government. The country had earlier experienced a complicated decade because of the power struggle between a monarchy backed by the United Kingdom and the United States and headed by the Shah Mohammad Reza Pahlavi and several other political groups, such as communists and democrats led by Mohammad Mosaddeq, so in the 1960s Iran was a socially effervescent country with multiple perspectives on the future of Iran, which then vanished because of the banning and censorship arising from the Revolution.

A brief historical explanation is required in order to properly understand that in the 1970s the Soviet government and the newly established theocracy had nothing in common ideologically, while

both were committed to using anti-western rhetoric in order to sidestep the context of strong democratic inclinations and (after the Bandung Declaration) to legitimizing their own struggle against the impositions of capitalists from the west, announcing that they would be acting on their own terms concerning international relations. The sociocultural features of Iran have implied that, in spite of the regime's views, there is an active digital arena in the country, i.e., while the theocracy ruled according to the Ayatollah Khomeini's conservative views, they seem to have adapted very well to innovations in the Internet sphere. Since the expansion of the Internet, the digital arena has been used by politicians, civil society and journalists, as well as religious elites who had been using digital aspects for theological debates [27].

At the end of the first decade of the 21st Century it was clear that this common position shared by both Russia and Iran was not only maintained but reinforced by the constantly developing technologies, such as the Internet and its evolving resources. Evgeniy Morozov, a Belarusian dissident familiar with the multiple strategies of the Soviet and post-Soviet regimes, acknowledged that technology is not unconditionally on the side of democracy, and more than a decade ago he announced that the cyberutopian conception was about to detonate in the hands of its adherents. This refers to the idea that the Internet, by virtue of its mere existence and the socialization it engendered, would find a way to becoming the main tool for democratization and open societies, and therefore defeat authoritarian regimes [28]. Morozov himself explained how this idealistic perspective failed, such as during the so-called Iranian "Green Revolution" or the Green Movement of 2009[1].

The ideas of Morozov have become even more relevant now that we face the twin challenges of cyber-attacks and AI, and with Iran and Russia also having become skilful and frequent users of these resources for furthering their political and geopolitical objectives. Coincidentally, according to the timeline presented by United States Institute for Peace, both countries started launching cyber-attacks around 2008 and 2009 [29]. Globally, the Cybersecurity and Infrastructure Security Agencies (CISA) registered a 38% increase in cyber-attacks in 2022 [30]. Although not all cyber-attacks come from the state actors themselves, here we are going to refer only to cyber-attacks undertaken by the governments of Russia and Iran. Both countries had been accused of drastically increasing cyber-attacks from Israel and the US in 2022 and 2023 [30]. Cyberwarfare from these perspectives is highly effective for both countries: since it has specific targets and because of its mechanisms it might be able

1 ——— Evgenii Morozov places the Green Movement or Green Revolution in Iran as one of the first collective and more illustrative movements greatly disappointed by the hopes of 'cyberutopianism'. Thousands of Iranians united in the streets of Tehran to speak out openly against the theocratic regime. Many of them organized the protest through Facebook, they even posted their precise location so others could join them in the street protests. The result, in terms of loosening the regime's tight grip, was a disaster. The political police traced the leaders of the Green Movement through geolocation, as many were subsequently arrested and harassed, including their families. In this case, the once supposedly liberating digital tools ended up aiding the persecution.

achieve its objectives with relatively low losses or no losses. Even so, this does not mean that this kind of attack is cheap to carry out.

4. Cyber-attacks on critical infrastructure: the strategy of Russia and Iran

The relevance of critical infrastructure for western countries' stability implies the analysis of cyber operations conducted by Russia and Iran. The low economic cost of the actions and the high impact in comparison with other kinetic attacks provides a justification for their use. In this section of the analysis, the research analysed the use of cyber operations in the framework of asymmetric actions, focusing on certain key operations. In the Russian case the analysis focused on the Ukrainian scenario, while Iranian operations demonstrate a higher diversity of targets and infrastructures affected.

4.1. Russia: hybrid warfare against Ukraine

The use of cyber strategies in Russia has been linked with informational warfare and influence operations. This is explained by the historical relevance of propaganda as a core element in political operations, a heritage from the Soviet times due to its long-term approach. Russia has extensively used trolls to manipulate, to create disinformation and to promote subversion. This use of cyber actions has been complementary to kinetic attacks against infrastructures which cause material damage, as happened in 2008 in Georgia and in 2014 in Ukraine [31]. The Russian invasion of Ukraine in 2022 has proved how hybrid warfare is a renewed, yet very aggressive way of attacking other countries' nervous systems, causing great damage at relatively little cost to the attacker. The very concept of hybrid war has been conceived of in terms of how Russia was able to find new vectors of attack – or what they call self-defence – since the annexation of Crimea in 2014:

> *The term Hybrid War or Warfare (HW) rose to prominence in defense and policy circles as well as in the media after the Russian annexation of Crimea in 2014. It was dragged out from the relative obscurity of military theory circles to become a mainstream term used to describe a myriad of seemingly different security and defense challenges to the West* [32].

Although the concept of Hybrid Warfare has been criticised it is still widely used, and it helps to explain several variations from the traditional conception of a physical war. The concept emerged first for

non-state actors who conducted operations with political or military objectives, then it also became part of the new military strategies for state actors. One of the main characteristics concerning Hybrid Warfare between states is the expansion of the battlefield:

> *In addition to blurred what is considered peace, conflict and war, hybrid warfare breaks the distinction between what is and what is not part of the battlefield... HW is both multimodal and employed on multiple levels at the same time, that comprises: the traditional levels of war – tactics, operation and field strategy- thereby accelerating the tempo at the strategic and tactical levels faster than a more conventional actor is able to do. Traditional physical spaces such as land, sea, air and space are increasingly accompanied by social and built spaces such as the political, economic, cultural and infrastructural and cyber* [32].

The concept of Hybrid Warfare refers not only to high-tech military capabilities and cyber weapons, but as Reichborn-Kiennerud and Cullen (2016) explain, the concept includes the cognitive and psychological factors also, which are key in achieving military objectives. Since the beginning of the Russian invasion of Ukraine in February 2022, there has been numerous significant cyber-attacks targeting Ukraine's energy sector. These attacks have had a significant impact on Ukraine's ability to generate and distribute electricity and have also caused widespread disruption to businesses and consumers. One of the most notable attacks was a distributed denial-of-service (DDoS) attack that targeted Ukraine's three largest electricity distribution companies in December 2021. The attack caused widespread outages, leaving millions of Ukrainians without power.

In February 2022, shortly after the start of the Russian invasion, Ukraine's national grid operator, Ukrenergo, was hit by a sophisticated cyber-attack that caused widespread power outages. The attack was attributed to Russia and was seen as a clear attempt to cripple Ukraine's infrastructure. In addition to the attacks on Ukraine's electricity grid, there have also been several attacks targeting Ukraine's oil and gas sector. In March 2022, a group of hackers calling themselves Killnet claimed responsibility for a cyber-attack that targeted Ukraine's state-owned oil and gas company, Naftogaz. The attack caused the company's website to go offline and disrupted its operations.

The cyber-attacks on Ukraine's energy sector have had a significant impact on the country's economy and have had a cumulative effect

when considering the consequences of the invasion. The attacks have caused billions of dollars in damage and have also led to a loss of confidence in Ukraine's energy sector. The attacks have also had a significant impact on the lives of ordinary Ukrainians, who have been forced to cope with power outages and other disruptions. The cyber-attacks on Ukraine's energy sector are part of a broader pattern of Russian aggression against Ukraine. The attacks are designed to weaken Ukraine's economy and infrastructure, and to make it more difficult for Ukraine to defend itself. The attacks are also a clear violation of international law and have been condemned by the United Nations and other international organisations.

The cyber-attacks on Ukraine's energy sector are a reminder of the growing threat of cyberwarfare. As the world becomes increasingly interconnected, cyber-attacks are becoming a common way for countries to wage war. The attacks on Ukraine are a wake-up call and highlight the need for countries to invest in cybersecurity and to develop strategies to deter and respond to cyber- attacks. However, western analysts say that many of the cyber-attacks inflicted by Russia against Ukraine have been quickly repaired, sometimes within hours, because of the highly skilled Ukrainian experts in these areas [33].

4.2. Iran: Cyber-attacks on critical infrastructure of western allies

Iranian cyber strategy is complementary to other influence operations in its areas of interest. Iranian geopolitics is based on generating deterrence by blocking the Hormuz strait or possessing ballistic missiles while deploying proxy actors on the ground, as happens in Syria, Lebanon, Yemen or Iraq. The strategy of using proxies has been employed in cyberspace also, which is a core feature of Iranian strategy [34]. Cyber capabilities have been extensively developed, thanks to the governmental cooperation with technological institutes and universities. In addition to scientific research, there are governmental investments in high-tech and communication companies. Most of these investments are direct from the Science Ministry, while others come from technological hubs [35].

Although we have already explained the key features underlying the tense relationship between Iran and the West over several decades, it is necessary to explain that Iran has conducted a long list of cyber-attacks since 2009, precisely when the radical anti-western president Mahmoud Ahmadinejad was elected. His aggressive rhetoric matched perfectly with the newly available tools at that time [28]. To

this purpose, Iran has developed both defensive capabilities against foreign aggression and against the regime's political rivals, as well as offensive capabilities to confront American superiority over digital infrastructures. The defensive capabilities are focused on protecting sensitive data and critical infrastructure against cyber-attacks. In the same manner, the Iranian government has developed measures against the coordination of anti-government groups, so as to prevent the introduction of western ideas in opposition to the regime. In contrast, offensive capabilities are developed as a complementary tool within an asymmetric strategy against their enemies [35].

Several analyses of tactics, techniques and procedures of Iranian cyberwarfare show similar patterns between the Iranian government and its proxies in the Middle East. This strategy has been widely employed since 1979, with Iran having a cohesive network in the region which also operates in the cyber domain. The network of actors is unstable and some of the organisations use similar resources, tactics and procedures. The similarities can imply confusion, with it being unclear as to who is behind the attacks or whether those responsible are acting under orders from the Iranian government, or whether proxies are acting independently with no direct instructions being given [36].

Iran's tense relationship with Israel has a long history, which starts in the religious and ideological terrain, but the conflict has escalated to political tensions and even overt threats in different periods since 1979. At the present time, the wide range of capabilities opened up by Hybrid Warfare has led Iran to commence an extensive sequence of operations within cyberspace to pursue objectives against countries perceived by the Iranian leadership as hostile. Iran has been – and still remains – a very active actor when it comes to cyber-attacks, and there are several groups that perform this kind of action. There have been a broad array of operations carried out since 2009, and one of the main strategies from Iran is to attack western allies in the Middle East, mainly Israel and Saudi Arabia [37][2].

One of the main attacks upon critical energy infrastructure was performed in 2012 against Saudi Aramco, "a company responsible for 10% of the world's oil supply at the time" [38]. This operation can be considered as industrial sabotage against the regional rival of Iran, which is a relevant ally of western countries [39]. The attack began on August 15, 2012, by means of malware called Shamoon, which began deleting and overwriting data in around 30,000 computers, and responsibility for this was claimed by a group called the Cutting Sword of Justice:

2 ——— For a complete timeline of Iranian cyber-attacks against different countries but mainly, United States, Israel and Saudi Araba see the USIP (May 3rd, 2023) report: Iran accelerates cyber-attacks. Available online https://iranprimer.usip.org/blog/2023/may/03/report-iran-accelerates-cyberattacks

> *The attacks were timed to coincide with Ramadan when most workers would be absent to allow the malware the maximum time to work unimpeded. The malware only infiltrated office computers and did not impact systems dealing with technical operations. Still, it grounded services to a halt, as office workers resorted to communications with typewriters and fax machines and gasoline refill trucks were turned away with no way to process payments. To mitigate the damage, Aramco purchased 50,000 hard drives, paying higher prices to cut the line and buy all the hard drives on the manufacturing line at several Southeast Asian factories* [38].

In the last decade, Iran has performed numerous cyber-attacks against several countries, mainly United States and its allies: continuous cyber-attacks against Israel, the United Kingdom, Australia and even Albania. An important feature of these actions is that they attack not only critical energy infrastructure but also infrastructure vital to health, as was the case in 2022 against Boston's Children's Hospital and on Israeli water facilities back in 2020 [37]. The theocratic government has also launched attacks during us elections.

> *The Office of the Director of National Intelligence said that it had "high confidence" that Supreme Leader Ayatollah Ali Khamenei authorized a cyber influence campaign during the 2020 presidential election. The online operation was intended to "undercut former President Trump's reelection prospects - though without directly promoting his rivals." Iranian cyber actors published more than 1,000 pieces of online content from several thousand fake social media accounts. Iran also sent threatening emails to Democratic voters, tried to exploit vulnerabilities on state election websites and attempted to hack the email accounts of political campaign officials* [37].

The attacks have continued during 2023 and will remain. In April 2023, Microsoft warned about the Iranian-linked group called Mint Sandstorm that has:

> *[...] started targeting critical U.S. infrastructure including energy companies, transit systems and seaports in 2021. The group gained access to sensitive systems "in support of retaliatory destructive cyberattacks [...] "The increased aggression of Iranian threat actors appeared to correlate with other moves by the Iranian regime under a new national security apparatus, suggesting such groups are less bounded in their operations."*

Regarding the complexity of the cyber-attacks from both countries, as a tool intended to destabilise or to act as a weapon (as done by Russia), it is expected that these attacks can and will be used widely in the future in many aspects, whether for criminal extortion, non-state actors and between States. International law has been left standing and it is highly unlikely that it could prevent this kind of action between states. The comparative case shows the relevance of the state as a core actor in cyberwarfare, which often sponsors non-state proxies as a means of avoiding attribution. In addition, it is relevant to consider the importance of public expenditure to improve cyber capabilities, as well as the coordination with scientific institutions and the private sector, which increase the complexity of the digital arena as a domain of the conflict.

4.3. The impact of cyber-attacks from Iran and Russia

The analysis and examples used to demonstrate that hybrid warfare is a widely used strategy for both Iran and Russia must also take into account the fact that its impact is somewhat ambiguous, just like the strategy itself. According to the literature review, a conventional conflict allows one to easily identify the main actors, their motivations and the consequences of their actions, while it is difficult to identify them in Gray Zone operations [11]. Our case study confirms that relations can be found between general strategies and specific actions, but due to the unconventional nature of the operations it is complicated to prove this entirely. As long as the perpetrators (groups of individuals) are possibly related to the regimes (both in Iran and Russia), they will probably remain as an important part of a clandestine or informal part of hybrid multimodal warfare.

In some cases, as happened in Georgia in 2008 or in Ukraine in 2014 [31], cyber operations were clearly used as a complementary tool for conventional Russian military actions. In those cases, cyberwarfare was a secondary means of supporting other types of operations having a defined authorship. Some of the Russian cyber operations could be included in the set of hybrid actions, as they had a connection with specific kinetic operations. Other actions, especially those related to energy infrastructure, would be more adequately classified within the Gray Zone spectrum, since they can condition further political negotiations [14]. In contrast, the Iranian operations in the cyber domain would be better classified as Gray Zone activities, as most of them were performed following political objectives to destabilise adversaries. Their actions would aim to generate deterrence in order to improve their geopolitical position [17]. According to this analysis, Iranian cyber operations involve a high number of

public and private actors [17], which increases the difficulty in clearly identifying the authors of the attacks.

As different modalities of hybrid or non-conventional operations can be easily tracked, as happens with proxy wars or some disinformation campaigns, cyberwarfare is even more obscure and difficult to analyse. Actors involved in cyber operations are multiple and not always directly linked with only one state, thus allowing for deflection of responsibility [20]. A further impact is that hybrid cyberwarfare has become a part of national geopolitical strategies and it will remain as such. It is important to acknowledge that while cyber-attacks are often initiated by Rogue States with authoritarian regimes, liberal western countries can indeed respond to these and fight back in the same ambiguous terms. When analysing cyberwarfare, there are immediate impacts from the actions involved, as happens with cyber-attacks against critical infrastructure, which are easy to identify. In contrast, it is even more difficult to fully prove the political long-term consequences of cyber operations conducted in the Gray Zone. Cyberwarfare as a tool for military operations produces clear effects in supporting kinetic actions, but those operations with geopolitical purposes are much more difficult to capture.

5. Conclusion

This research constitutes an initial approach to the use of cyberwarfare against targets belonging to the energy sector. In a hyperconnected globalized world, various kinds of critical infrastructure are vulnerable to cyber-attacks. The article is intended to present a comparative analysis of the use of cyberwarfare by Russia and Iran. These cases show how two rogue states have included cyber actions as an important tool within their general strategy based on asymmetric operations. As is evident from our analysis, the actions implemented at the operational level are perfectly coordinated, combining state and non-state actors and having a long-term approach of weakening their adversaries.

Their strategies include cyber actions in the framework of hybrid warfare. Despite this concept having been widely brought into question, it is still used in official speeches and analysis [32]. The cyber tools are inserted into the framework of Gray Zone conflicts, as their use can weaken the defences of adversaries. The consequences of cyber actions can imply cognitive and psychological victories which can increase the complexity of the adversaries' social environment. The case of Ukraine provides evidence for some of the direct effects

of cyber-attacks on energy infrastructure, such as blackouts and the interruption of normal business activities. Such actions have been mainly based on service denial, implying both material and reputational damage. The analysis has shown that since the beginning of the invasion of Ukraine, Russia has developed several cyber actions within the framework of a general strategy. The purpose of such actions has been to support conventional military operations, as well as to weaken an adversary's defence system and undermine the morale of its citizens.

In addition to Russia, Iran has conducted cyber-attacks over a long period of time. This fact shows the long-term approach of their strategy. In this research we have analysed various actions taken against oil infrastructure in Saudi Arabia, American healthcare facilities and water supply in Israel. In the Iranian case we can see a high diversity of targets across different countries, but at the same time they use cyberwarfare to complement other offensive and soft power strategies. The Iranian case is an interesting one to study, as the regime combines high-tech elements in the digital arena with an ideological structure established within the cognitive framework of the regime.

This article facilitates the exploration of future research avenues for conducting deeper examinations into operations carried out in the Gray Zone. As the cyber dimension is a core element in the strategies of certain states, we cannot ignore the relevance of social and human dimensions for understanding the full impact of cyberwarfare against adversaries. From an analytical perspective, it would be relevant to have greater knowledge of western cyber operations conducted against rogue states, in order to find parallels in their procedures. Other future lines of research could be focused on understanding the various effects related to deterrence aspects provided by cyber capabilities, as well as the various societal consequences arising from energy infrastructures being attacked. These lines of research could be strengthened by producing primary data through interviews with cyber experts to find weaknesses and strengths in the current energy systems. In the same way, it would be relevant to produce quantitative data regarding social perceptions about the consequences of a lack of energy supply on society, seeking to study social resilience in Western countries.

In addition, it is important to acknowledge the limitations of this paper. First, it is complex to analyse the phenomena of cyberwarfare, owing to its particular characteristics: the blurred attribution of responsibilities; the lack of internet regulation and law enforcement within cyberspace. Another important limitation is the feature of

non-state proxies linked to cyber-attacks to strengthen a state's political or economic objectives. It is equally difficult to measure the effectiveness of such attacks for achieving Iran and Russia's geopolitical goals. The analysis presented here clearly shows that the cyber-attacks are destabilising energy infrastructure, while legal loopholes and poor law enforcement, in conjunction with the ambiguous nature of the attacks themselves, makes the potential damage incurred difficult to acknowledge or confront.

References

[1] A. Pinedo Lapeña, "Ciberseguridad, geopolítica y energía," in *Energía y Geoestrategia*, Spanish Ministry of Defense, Madrid: Spanish Institute for Strategic Studies, 2022, pp. 159–196.

[2] K. Melligan, "The Vulnerability of the United States Electrical Power Grid," *Journal of Applied Business and Economics*, vol. 22, no. 7, pp. 155–163, 2020, doi: 10.33423/jabe.v22i7.3259.

[3] Z. Zhang, "Cybersecurity policy for the electricity sector: the first step to protecting our critical infrastructure from cyber threats," *Boston University Journal of Science & Technology Law*, vol. 19, no. 2, pp. 319–366, 2013.

[4] A. Yates, "Death modes from a loss of energy infraestructure continuity in a community setting," *Homeland Security & Emergency Management*, vol. 10, no. 2, pp. 587–608, 2013, doi: 10.1515/jhsem-2012–0048.

[5] E. Hatipoglu, S. Al Muhanna, B. Efird, "Renewables and the future of geopolitics: Revisiting main concepts of international relations from the lens of renewables," *Russian Journal of Economics*, vol. 6, no. 4, 2020, pp. 358–373, 2020, doi: 10.32609/j.ruje.6.55450.

[6] J.A. Lewis, "The Electrical Grid as a Target for Cyber Attack," 2010. [Online]. Available: http://csis-website-prod.s3.amazonaws.com/s3fs-public/legacy_files/files/publication/100322_ElectricalGridAsATargetforCyberAttack.pdf [Accessed: Dec. 29, 2023].

[7] A. I. Ayerbe, *La ciberseguridad en el sector energético*, ARI 3/2020. Madrid: Real Instituto Elcano, 2020.

[8] J.S. Nye, "Soft power," *Foreign Policy*, vol. 80, pp. 153–171, 1990.

[9] J. S. Nye, *Soft power: The means to success in world politics*. New York: Public Affairs, 2004.

[10] D. Belo, "Conflict in absence of war: a comparative analysis of China and Russia engagement in gray zone conflicts," *Canadian Foreign Policy Journal*, vol. 26, no. 1, pp. 73–91, 2020, doi: 10.1080/11926422.2019.1644358.

[11] D. Ventre, *Cyberwar and Information Warfare*. London: ISTE, 2011.

[12] D. Yergin, "Ensuring Energy Security," *Foreign Affairs*, vol. 85, no. 2, pp. 69–82, 2006, doi: 10.2307/20031912.

[13] A. Sánchez-Ortega, *Poder y seguridad energética en las relaciones internacionales: la estrategia rusa de poder*. Granada: Editorial Universidad de Granada, 2012.

[14] S. Paltsev, "The complicated geopolitics of renewable energy," *Bulletin of the Atomic Scientists*, vol. 72, no. 6, pp. 390–395, 2016, doi: 10.1080/00963402.2016.1240476.

[15] G. Escribano, "Geopolítica de la energía: identificación de algunas variables," *Índice: Revista de Estadística y Sociedad*, vol. 46, pp. 12–14, 2011.

[16] J. Jordán, "Un modelo de análisis geopolítico para el estudio de las relaciones internacionales," Documento *Marco 04/2018, Instituto Español de Estudios Estratégicos*, 2018.

[17] J. J. Wirtz, "Life in the "Gray Zone": observations for contemporary strategists," *Defense & Security Analysis*, vol. 33, no. 2, pp. 106–114, 2017, doi: 10.1080/14751798.2017.1310702.

[18] J. W. Matisek, "Shades of Gray Deterrence: Issues of fighting in the Gray Zone," *Journal or Strategic Security*, vol. 10, no. 3, pp. 1–26, 2017.

[19] M. J. Mazarr, *Gray Zone: Understanding a Changing Era of Conflict*. Carlisle: United States Army War College Press, 2015.

[20] R. Stiennon, *Surviving cyberwar*. Plymouth: Government Institutes, 2010.

[21] E. Schmidt, J. Cohen, T*he new digital era: Reshaping the future of people, nations and business*. New York: Random House, 2013.

[22] A. Greenberg, *Sandworm: A new era of Cyberwar and the Hunt for the Kremlin's most dangerous hackers*. New York: Doubleday, 2018.

[23] B. Groys, "Russia and the West: The Quest for Russian National identity," *Studies in Soviet Thought*, vol. 43, no. 3, pp. 185–198, 1992.

[24] J. Mankoff, "Russia and the West: Taking the longer view," *The Washington Quarterly*, vol. 20, no. 2, pp. 123–135, 2007.

[25] E. Rummer, A. Stent, "Russia and the West," *Survival: Global Politics and Strategy*, vol. 51, no. 2, pp. 91–104, 2009, doi: 10.1080/00396330902860835.

[26] M. Skladanowski, "The Myth of Russian Exceptionalism: Russia as a Civilization and its Uniqueness in Aleksandr G. Dugin's Thought," *Politics, Religion and Ideology*, vol. 4, no. 20, pp. 423–446, 2019, doi: 10.1080/21567689.2019.1697870.

[27] N. Mina, *Blogs, cyber-literature and virtual culture in Iran*. George C. Marshall: European Center for Security Studies, 15, 2007.

[28] E. Morozov, The net delusion. *The Darkside of the Internet Freedom*. New York: Public Affairs, 2012.

[29] United States Institute for Peace, "Report. Iran accelerates cyber-attacks," May 3, 2023. [Online]. Available: https://iranprimer.usip.org/blog/2023/may/03/report-iran-accelerates-cyberattacks. [Accessed: Dec. 29, 2023].

[30] Check Point Research, "Check Point Research reports a 38% increase in 2022 global cyberattacks," Jan. 5, 2023. [Online]. Available: https://blog.checkpoint.com/2023/01/05/38-increase-in-2022-global-cyberattacks/. [Accessed: Jan. 25, 2024].

[31] T. Maurer, G. Hinck, "Russia: Information Security meets cyber security," in Confronting *an axis of cyber? China, Iran, North Korea, Russia in Cyberspace*, F. Rugge, Ed. Milan: Institute for International Political Studies (ISPI), 2018, pp. 39–57.

[32] E. Reichborn-Kjennerud and P. Cullen, "What is hybrid warfare?" 2016. [Online]. Available: https://www.jstor.org/stable/pdf/resrep07978.pdf. [Accessed: Dec. 29, 2023].

[33] The Economist, "Lessons from Russia's cyberwar in Ukraine," Science and Technology, Nov. 30, 2022. [Online]. Available: https://www.economist.com/science-and-technology/2022/11/30/lessons-from-russias-cyber-war-in-ukraine. [Accessed: Dec. 29, 2023].

[34] L. Tabanski, "Iran's cybered warfare meets western cyber-insecurity," in *Confronting an axis of cyber? China, Iran, North Korea, Russia in Cyberspace*, F. Rugge, Ed. Milan: Institute for International Political Studies (ISPI), 2018, pp. 121–141.

[35] G. Siboni, S. Kronenfeld, "Iran and Cyberspace Warfare," *Military and Strategic Affairs*, vol. 4, no. 3, pp. 77–99, 2012.

[36] J. G. Spataro, Iranian cyber espionage. Master Thesis. Utica College, 2019.

[37] United States Institute for Peace, "Report. Iran accelerates cyber-attacks," May 3, 2023. [Online]. Available: https://iranprimer.usip.org/blog/2023/may/03/report-iran-accelerates-cyberattacks. [Accessed: Dec. 29, 2023].

[38] United Against Iranian Nuclear, "Report: The Iranian Cyberthreat," UAIN, 2023. [Online]. Available: https://www.unitedagainstnucleariran.com/history-of-iranian-cyber-attacks-and-incidents. [Accessed: Dec. 29, 2023].

[39] S. Jones, "Cyber warfare: Iran opens a new front," Financial Times, Apr. 26, 2016.

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

The Russia-Ukraine Conflict from 2014 to 2023 and the Significance of a Strategic Victory in Cyberspace

Dominika Dziwisz | Jagiellonian University, ORCID: 0000-0002-5837-3446
Błażej Sajduk | Jagiellonian University, ORCID: 0000-0002-2974-8173

Abstract

The article explores Russian engagement in cyberspace during the conflict with Ukraine. Many experts have been surprised not only by the lack of coordination between offensive military operations in cyberspace and other domains, but also by the absence of significant cyberattacks. The central argument revolves around the perceived inadequacy of Russian capabilities. However, the authors contend that such an assessment is flawed and stems from the imposition of Western expectations onto a non-Western actor. They argue that the Russians' employment of cyberspace not only aligns with their strategic culture but also represents a continuation of their utilisation of cyber as a tool for disinformation, which was previously observed during the war with Georgia in 2008 and the initial phase of the conflict with Ukraine in 2014. The aim of the article is threefold. Firstly, it discusses the Western strategic discourse regarding the potential use of cyberspace in warfare. In contrast to the position of Western experts, the second part of the article presents the Russian approach. The third section describes how the application of Russian cyber warfare concepts has played out in practice during the conflict in Ukraine.

Keywords
Cyberspace, Russia, Ukraine, war, strategy

Received: 23.08.2023

Accepted: 5.12.2023

Published: 28.12.2023

Cite this article as:
D. Dziwisz, B. Sajduk "The Russia-Ukraine conflict from 2014 to 2023 and the significance of a strategic victory in cyberspace," ACIG, vol. 2, no. 1, 2023, DOI: 10.60097/ACIG/162842.

Corresponding author:
D. Dziwisz, Jagiellonian University; ORCID: 0000-0002-5837-3446; E-MAIL: dominika.dziwisz@uj.edu.pl

1. Introduction

During the 2013 gathering of high-ranking Russian and US defence officials, General Nikolai Makarov derided the absence of information warfare in the mission of US Cyber Command (USCYBERCOM) [1]. In a bold speech, he told his counterparts, "One uses information to destroy nations, not networks" and suggested that the Americans' lack of emphasis on information warfare demonstrated their ignorance. This incident served as a clear indication of Russia's cyberspace priorities, as subsequently reflected in their strategic documents and implemented during the Ukraine conflict in 2022.

Despite the ongoing war in Ukraine, significant breakthroughs on the battlefield resulting from cyberattacks have yet to materialise. During the Russian-Ukrainian conflict, many experts have expressed surprise at the lack of offensive cyber actions. However, an analysis of cyberattacks since 2014 indicates that the Russians never considered cyberspace as a decisive domain for offensive actions [2], [3], [4]. From 2000 to 2020, Russia primarily focused on intelligence activities. Approximately 61% of attributed incidents were centred on the acquisition of information rather than disruption or degradation of adversary systems [5]. Furthermore, coordination between cyber operations and military actions has not unfolded as expected. In contrast to initial attempts to synchronise cyber and kinetic forces at the beginning of the war, we now observe the independent use of these two Russian capabilities [2]. This discrepancy may be attributed to the different objectives assigned to Russian cyber operations and kinetic invasions. Cyber operations focus on information warfare, including disinformation, propaganda, and subversion, while kinetic actions aim to acquire territory. As a result, it can be deduced that the highly anticipated "cyber Pearl Harbor" event is unlikely, and Russia's performance in cyber warfare is not worse than expected. This is primarily due to the fact that cyber weapons are not suited to circumstances in Ukraine.

The article reviews opinions regarding the role of cyberspace in Russian strategy. Attention was drawn to the divergent understanding among Western experts regarding the strategic utilisation of cyberspace by the Russians. This discrepancy contradicted both earlier assessments and actual Russian actions, starting from the attacks on Estonia in 2007.

Consequently, the following research hypotheses have been adopted:

H1: A different understanding of the use of cyberspace for strategic purposes, compared to the Russian perspective,

led to the formation of numerous inaccurate expectations and forecasts regarding cyberspace use during the war in Ukraine.

H2: Cyberspace did not effectively serve Russia's objective of territorial acquisition in Ukraine, because it is better suited as a domain for operating in the grey zone, specifically for informational purposes.

H3: Cyberspace was mostly utilised by the Russians in the initial phase of the war to deploy offensive weapons against Ukrainian command and control systems, as well as massive malware attacks.

To investigate these hypotheses, the authors conducted a detailed analysis of assumptions and predictions on significance of cyberspace use for strategic objectives.

To conduct the study, a registry and database were developed, containing scientific articles, public writings, as well as reports from official think tanks and governments concerning the strategic use of cyberspace by the Russian Federation, with particular emphasis on publications related to cyber activities accompanying the conflicts in 2014 and 2022. Based on this, a study of source material was carried out using a critical analysis method.

This paper will proceed as follows: The first section describes Western perceptions of cyberspace use during conflicts, starting from the cyber Pearl Harbor and ending with actions below the threshold of war. The second section discusses the Russian strategic discourse on the role of cyberspace during conflict and warfare. The third part deals with the issue of Russian offensive actions in cyberspace and their role in achieving strategic victory.

2. Western Strategic Discourse: From Cyber Pearl Harbor to the Cyber Grey Zone

War is a legally and morally exceptional state of affairs, well defined on the grounds of international law. However, predictions about the future of war follow narratives and intellectual trends. Various manifestations of war, e.g. hybrid war, cyberwar, grey zone confrontation, come to the forefront of academic debate when social circumstances become favourable. Moreover, the development of cyberwar-related topics has resulted in a division within the field

between "alarmists" who view cyber power as crucial in modern strategic affairs and "sceptics" who believe that cyber power possesses less potency. The multitude of views regarding the potential use of cyberspace in warfare, as well as the ambiguity surrounding the terminology employed, may lead to, among others, a misunderstanding of Russian operational concepts.

The warning issued by US Defence Secretary Leon E. Panetta in 2012 about an unavoidable "cyber Pearl Harbor", an attack that would cause physical destruction and loss of life, influenced the understanding of conflicts in the digital realm, where the sole alternative to cyberwar is cyberpeace [1]. Since then, "exaggeration" has become an important characteristic of the cyberwar discourse (for example, exaggerating the effect of cyberattacks on Estonia in 2007 or the Russian invasion of Georgia in 2008) [6]. This concept found fertile ground, especially among high-ranking US military officials, particularly as a means to rationalise heightened investment in cybersecurity. In an unclassified memorandum dated 23 March 2012, General Keith Alexander provided a strategic assessment for operating in cyberspace and "Preventing a Pearl Harbor Environment" [7]. He shared his viewpoint on the potential occurrence of a cyber "Pearl Harbor" and delved into the perils associated with failures in the realm of cyberspace. This analogy and metaphor quickly caught on, not only in official speeches by government officials but also in media coverage, where they were uncritically repeated. It also heavily influenced the global discourse on cybersecurity and strategic planning in the early 2000s [8]. However, this circumstance was not without adverse repercussions. The ease of using catchy metaphors in discussions about war encouraged the unquestioned expansion of a reasoning that appears effective in theory but lacks explanatory capability in practice. Those who overlook this tendency are prone to rely on metaphors to do their thinking for them [9].

The widespread adoption of terminology such as "cyber-doom", "power grid shutdown", "shock and awe", and "worst-case scenarios" also garnered support from some researchers, particularly leading up to and during the onset of the Russian-Ukrainian war. Jason Healey, the former Director of the Atlantic Council's Cyber Statecraft Initiative, predicted that "it will be the first time a state with real capabilities is willing to take risks and put it all on the line" [10], and that "a Russian cyber offensive might have far more impact on the battlefield, more coercive power, more lethal and widespread effect than many doubters would expect" [11]. William Courtney and Peter A. Wilson of the RAND Corporation wrote that a Russian invasion would "likely employ massive cyber and electronic warfare tools and long-range

PGMs to create 'shock and awe,' [and] causing Ukraine's defences or will to fight to collapse" [12]. Keir Giles of Chatham House believes that "a destructive cyber onslaught could target military command and control systems or civilian critical infrastructure and pressure Kyiv into concessions and its friends abroad into meeting Russia's demands" [13]. NATO analysts David Cattler and Daniel Black assert that "cyber-operations have been Russia's biggest military success to date in the war in Ukraine" [14]. Despite some limitations, Russian cyberattacks on Ukrainian government and military command centres, logistics, emergency services, and crucial facilities such as border control stations were completely aligned with a strategy known as "thunder run", aimed at generating chaos, confusion, and uncertainty, and ultimately to prevent a costly and prolonged war in Ukraine. It is worth noting that Russian cyber-units have showcased their capability to achieve success with minimal prior warning and guidance, despite the significant challenges impeding Russia's military endeavours [14].

Despite these radical predictions, cyber operations don't appear to be playing a decisive role on the Russian-Ukrainian battlefield. Since the beginning of the war, various, sometimes contradictory, analyses have been published regarding use of the cyberspace in this conflict. However, most experts agree on one aspect – cyber operations did not significantly contribute to achieving Moscow's campaign objectives. James Lewis from CSIS writes that "the so-far inept Russian invasion, where cyber operations have provided little benefit, raises questions about the balance between defence and offense in cyberspace, the utility of offensive cyber operations, and the requirements for planning and coordination" [3]. Jon Bateman from the Carnegie Endowment for International Peace states that "Russia's cyber operations in Ukraine have apparently not had much military impact", and even goes so far to describe it as "Russia's humbling experience" [15]. On the other hand, John Hultquist from Mandiant points out that "many of these attacks carried out were designed to affect the civilian populace rather than any military targets" [16]. Marcus Willetta from IISS was surprised that "Russia's invasion of Ukraine in 2022 did not appear to be accompanied from the outset by Russian cyber operations aimed at extensively disabling Ukraine's critical national infrastructure" [4].

Microsoft wrote about the "limited impact" of cyber operations and the sharp decline in their intensity and pace already at the beginning of March 2022 [4]. Researchers Nadiya Kostyuk and Erik Gartzke say that "while Russia has conducted some cyber operations in Ukraine, both in the lead-up to and after the February invasion, these have

neither supplanted nor significantly supplemented conventional combat activities" [2].

There are several factors that can explain the lack of spectacular successes by the Russians in cyberspace, including a lack of flexibility in army management, the desire to avoid risks associated with the uncontrolled spread of attacks to other countries, the plan for a swift victory in the early weeks of the war without the need to utilise cyber capabilities [18], as well as the lack of coordination between cyber and kinetic operations [2]. There are also voices suggesting that Russian military strategists set the bar too high for cyber operations, basing their planning on observations from wars fought in the 1990s and the beginning of the current century, without adapting them to the conditions of total war [19]. There was a lack of ideas (and possibly processing power or capability) for coordinating actions across different domains of warfare. Despite attempts in the early weeks of the invasion, currently, we can only observe independent utilisation of Russian capabilities [17].

However, another explanation for the absence of a cyber Pearl Harbor cannot be ruled out. Namely, that from the very beginning, the Russians did not plan for wide-scale use of direct cyber capabilities against critical infrastructure objects, not due to a lack of such capabilities, but rather because of other strategic assumptions that perceive the cyberspace as most useful for achieving informational objectives. If this is the case, Russia may have different strategic goals for the use of cyberspace. This also fits into the current decline in popularity of the term "cyberwar", as multiple non-military perspectives on understanding cyberpower are emerging. A review of the state of the art has shown that competition below the threshold of armed aggression is constantly gaining in importance. The emphasis on activities in the grey zone appears in, e.g. strategic documents of the largest cyber rivals – the US, Russia and China – but also in national security strategies of other countries, including Australia, Germany, Great Britain, and Indonesia [20]. The most contemporary approach perceives cyberpower mostly as a form of intelligence activity [21] and cyberpower exercises as a state of "unpeace" [22], an equivalent of the terms: "grey zone" between war and peace [23], [24] (the most popular), "non-war military activities" [25],"warfare during peacetime" [26], [27], "subliminal aggression", "persistent cyberspace confrontation", or "non-war" [20], [28]. All these terms refer to actions below the threshold of armed aggression and usually cover the entire spectrum of possible actions, not only those in cyberspace. Therefore, besides deriving offensive and defensive strategies from the study of war, in practice, cyber conflict has been low in intensity, remaining below the threshold of armed conflict [21].

However, the ongoing war confirms that the unquestionable benefits of cyber operations during a conflict below the threshold of war lose significance when the conflict becomes "hot". The key advantage of cyber actions, or attribution – the clear indication of the attacking entity in cyberspace – loses significance when both sides are already in physical confrontation, and their mutual intentions are clear. In other words, deniability and ambiguity, which define grey zone conflicts, do not apply during times of war.

One of the advantages of conducting hostile operations in cyberspace is the ease of disrupting enemy information exchange, which can be more effectively achieved, for example, through missile attacks on telecommunication infrastructure elements. The third advantage is their non-territorial nature, meaning they can be carried out from any location on Earth, but this loses significance when kinetic targets can be attacked throughout the enemy's territory, as the Russians are doing by targeting objectives across Ukraine. In the current phase of the war, Russia continues to utilise cyberspace to conduct operations in the grey zone against states supporting Ukraine. As a result, one can expect an intensification of disinformation and intelligence activities. This is reflected in opinions from Microsoft experts, who indicate that hostile Russian actions aimed at states supporting Ukraine primarily have an intelligence character. For instance, the attacks targeting Polish entities were not intended to damage systems as much as to gather information about the logistics process related to providing assistance to Ukraine.

Despite the aforementioned factors, which prevent categorising current cyber activities of the war as "grey zone" actions, the techniques employed in Ukraine remain similar to those utilised prior to 24 February 2022, once the element of surprise is excluded.

3. Russian Strategic Discourse – Information as a Weapon

To comprehensively grasp the broader context of Russian activities in cyberspace during the war with Ukraine, it becomes imperative to delve into how Russia defines and assigns significance to these activities at a strategic level. Undoubtedly, perception of this role is influenced by a longstanding tradition rooted in the development of doctrines pertaining to the active utilisation of intelligence and subversive operations, tracing back to the eras of Tsarist Russia and the Soviet Union. Russia's modern armed forces exhibit a creative continuity of this tradition. The very notion of "information

warfare" can be viewed as a natural extension of concepts formulated in the 1920s regarding active intelligence and counterintelligence. As posited by Jolanda Darczewska, this concept signifies "not so much a change in the theory of its conduct (the changes mainly relate to the form of its description, and not the content), but rather a clinging to old methods (sabotage, diversionary tactics, disinformation, state terror, manipulation, aggressive propaganda, exploiting the potential for protest among the local population)" [29].

Historical heritage played a significantly larger role in contemporary Russian military strategic thought. This is because it is influenced by two conflicting perspectives. According to Dimitri Minic, on one hand, it is shaped by arguments advocating the traditional definition of war as "the direct and open use of armed violence". The opposing view posits that the central issue is the "bypassing of armed struggle" through the use of "indirect, non-armed violence", including activities in the cyber sphere [30]. This duality in defining the role of non-kinetic actions conducted in cyberspace (as well as the infosphere) at the strategic level may explain the limited role of cyber offensive actions during the hot phase of the conflict with Ukraine.

In addition to considering the historical context and distinctive strategic culture, the Russian approach to information and its role in achieving objectives within international politics and internal security is shaped and refined through numerous official documents [31]. These documents unequivocally indicate Russia's awareness of being perceived as a threat by numerous countries. Concurrently, Russia is cognizant of its relatively disadvantaged position in the event of a confrontation with NATO. This is particularly evident in the 2021 National Security Strategy of the Russian Federation, wherein explicit mention is made of foreign global internet companies that disseminate disinformation and orchestrate social protests based on "the objective social and economic difficulties in the Russian Federation" [32]. Moreover, the Russian strategic culture perpetually portrays Russia as a besieged fortress [33], with the country's power elite steadfastly believing that it faces an incessant threat of cyberattacks from the West, particularly NATO [34].

These factors create a foundation for the underlying assumptions of the Russian strategy in global competition, wherein continuous competition in the information domain is viewed as a permanent aspect of Russia's exertion of pressure on Western states [35]. As a comparatively weaker actor, Russia must maintain a persistent and proactive approach in influencing other countries. This strategic outlook is operationalised at the military level through a collection

of concepts known as "Gerasimov's Doctrine" [36], which was largely a reaction to US offensive actions, according to Moscow [37]. A crucial component of this doctrine is the belief in the necessity of conducting "active defence", which entails employing non-military means and indirect approaches to maintain constant pressure on adversaries [38]. The concept of "active defence" encompasses a wide array of activities aimed at systematically destabilising the social, political, and military systems of the opponent over an extended period, preceding any kinetic actions. Key elements for exerting this pressure involve non-military means utilised below the threshold of war, such as psychological warfare and subversion.

When discussing the evolution of warfare, Russian sources indicate that the current sixth generation of warfare involves "high-precision weapons based on land-air-sea", with cyberspace assuming a reduced role as "informational-space support" [38]. The Russians classify information warfare activities into two interconnected and complementary categories: information operations and cyber operations (i.e. offensive operations in cyberspace as defined by NATO) [39]. The latter further encompasses two distinct strands: cyber-psychological and cyber-technical operations. Cyber-psychological operations primarily leverage platforms, such as social media, to disseminate disinformation and propaganda, intending to exert long-term influence on societies and potentially destabilise hostile states. On the other hand, cyber-technical operations include a broad range of activities targeting enemy infrastructure. In the Western paradigm, however, greater emphasis is placed on destructive offensive cyber operations targeting critical infrastructure, rather than information operations [40]. It is essential to note that Russian military terminology distinguishes their approach to information warfare, which extends beyond activities conducted solely during or immediately preceding kinetic warfare, in contrast to the Western approach that focuses on the tactical utilisation of information warfare during ongoing conflicts [41]. In this regard, Keir Giles astutely observed that the Russian term "kibervoyna" (cyber war) is only used when referencing Western thinking rather than Russian approaches [42].

During a conflict, Russia focuses on enhancing its armed forces to conduct strikes against critical infrastructure. However, the primary role of this task falls under long-range strike capabilities, specifically cruise and ballistic missiles, with cyber capabilities providing supporting roles [43]. It is worth noting that "It remains unclear how cyber weapons fit into Russian thinking on strategic operations and SODCIT (*Strategic Operation for the Destruction of Critically Important Targets*) in particular" [43]. Despite this, artillery remains a significant

component in Russia's "non-contact warfare" [38] approach, relegating cyberspace to a secondary position. According to a report from the RAND Corporation, "Russian military officers and analysts believe that augmenting capabilities in EW, space, and cyber could fully compensate for a lack of conventional theatre strike capacity" [44]. In the Russian armed forces, cyberspace is not regarded as a novel weapon category that fundamentally alters the nature of temporal activities on the battlefield. Instead, it is viewed as a tool primarily for subversion and enhancing its effectiveness. This perspective has guided the approach of the Russian Federation's armed forces in recent years.

4. Russia's Utilisation of Cyberspace During an Armed Conflict

When examining the utilisation of cyberspace in warfare, a crucial aspect pertains to its application during military interventions conducted by Russia against neighbouring countries. The Russian power elite justified these interventions as defensive actions aimed at safeguarding Russia through what they perceived as limited-scale defensive wars [45]. In this regard, the actions taken against Estonia in 2007 are particularly important, but did not cross the threshold of physical interference by armed forces. Additionally, the armed conflicts with Georgia in 2008 and Ukraine in 2014 exemplify Russia's approach.

In Estonia, the pressure exerted was primarily achieved through successful yet temporary distributed denial-of-service (DDoS) attacks targeting government IT systems. However, no substantial cyberattacks have been officially confirmed, and experts have noted the absence of such attacks in Moscow's arsenal. During the war with Georgia, cyber activities were predominantly ancillary to kinetic operations. Similar to the cyberattack on Estonia, instances of DDoS attacks and website defacements against official institutions were reported. Nonetheless, the Georgia conflict in 2008 demonstrated that offensive operations in cyberspace need not occur at the "speed of cyber" [46]. The coordination of such operations with other domains poses a challenge that is difficult for most armed forces worldwide to manage. In the context of the Georgia conflict, Erik Gartzke astutely noted that Russia relied on conventional forces rather than cybernetic forces to achieve success [47].

In a similar vein, during the initial phase of the conflict with Ukraine in 2014, the utilisation of offensive actions in cyberspace did not

hold significant importance in warfare [48]. Researchers and analysts posed the question: "Why was there no cyberwar in Ukraine?" [49]. James A. Lewis, when evaluating Russian offensive activities in cyberspace targeting Ukraine in 2015, observed "Cyberattacks are a support weapon and will shape the battlefield, but by themselves they will not produce victory" [50]. Subsequent cyberattacks on the Ukrainian power grid in 2015 and 2016 were primarily employed to exert pressure on Ukrainian society and the government in Kiev [51]. The NotPetya attack in 2017 aligns with the same logic of activities in the grey zone. It is important to emphasise that none of these actions changed the course pursued by authorities in Tallinn, Tbilisi, or Kiev. This fact certainly did not escape the attention of the Kremlin's ruling elite. Hence, it appears that Russian expectations regarding activities in cyberspace are much more modest than assumed in the West, a notion seemingly substantiated by the progression of the Russian invasion of Ukraine in 2022. This is consistent with the conclusions drawn by analysts at CSIS in 2023: "Moscow appears to view using cyber operations more as a means of harassing Ukraine and supporting information operations than as a war-winning weapon indicative of the thunder run strategy (...) Cyber operations remain a weak coercive instrument for Moscow despite their frequent use" [52].

The shift by Russia from operations in the grey zone to a kinetic military operation can be explained not only by the ineffectiveness of such actions but also by Russia's increased assertiveness in international relations over the past decade and Vladimir Putin's growing acceptance of higher risk levels, particularly in actions directed against Russia's immediate surroundings [53]. Additionally, Tor Bukkvoll highlights that Putin's willingness to take on more risk stems from the "prospect theory", which posits that individuals who fear losses are more inclined to engage in risky actions compared to those pursuing profit [54]. Consequently, it can be assumed that the fear of conflict escalation did not constrain Russian activities in cyberspace, and if Russia possessed effective cyber weapons, they would have undoubtedly been employed already. The level of tactical planning is evident in Russian actions, as the dominant attack tools were modified and gradually adapted in preparation for the impending invasion. Kenneth Geers noted in this regard that the beginning of 2022 witnessed a prevalence of defacement attacks, followed by intensified distributed denial-of-service (DDoS) attacks just before the invasion, and massive-scale malware usage during the kinetic phase of the operation [55].

Equally important in understanding the role of cyberspace activities in kinetic conflicts is Russia's extensive employment of malware.

Microsoft has identified at least eight families of malware utilised in the initial weeks of the attack [56]. However, determining the extent of their coordination with kinetic assaults poses challenges [15]. This aligns with the fact that cyberspace activities are subject to limitations that require a choice between mutually exclusive attributes such as speed, intensity, or control. Lennart Maschmeyer referred to this predicament as the subversive trilemma [57]. It appears that Russia, in this trilemma, prioritised intensity at the expense of the other factors, recognising that leaving the grey zone would hinder their ability to maintain coordinated speed between cyberspace activities and operations in other domains. Similarly, they relinquished the control component. Nonetheless, these limitations restrict the ability of cyber operations to successfully produce independent strategic utility. Herbert Lin suggests that a potential solution could involve increasing the scale of cyberattacks at the expense of quality, selecting tactics that "go forth and damage Ukrainian institutions that provide government, military, and economic functions, that inform the Ukrainian public, or that constitute Ukrainian critical infrastructure" [58]. However, this approach has its limitations, as the Russians were unable to sustain the same intensity after the initial phase of cyberattacks from January to April 2022 [59]. The offensive role of cyberspace activities was likely constrained, partly because the Russians focused on psychological impact and information warfare, inadvertently exposing their covert access to Ukrainian IT systems, which could have adverse consequences for future offensive cyber operations. This suggests that this strategy might make it impossible to reue vulnerabilities and accesses gained during grey zone operations in a full-scale war, as the adversary may update their systems and bolster defences [61]. However, this may indicate a deliberate Russian prioritisation of grey zone conflict characteristics in cyberspace. The extensive use of malware resulted in some targets being infected with both malware and subjected to kinetic attacks, which could create the illusion of a partial correlation between offensive cyberspace and kinetic actions. This raised doubts among certain Western experts [17]. Nevertheless, it is evident that even if highly coordinated, the impact of cyberspace activities on the overall course of the war has thus far been limited. Despite the increased number of Russian cyberattacks in the initial phase, most proved unsuccessful: "only 29 percent of the attacks breached the targeted networks – in Ukraine, the United States, Poland and the Baltic nations (...) only a quarter of those resulted in data being stolen" [62].

It is noteworthy that the Russians did not show significant interest in synchronising their state-of-the-art electronic warfare systems with

other types of weapons. Jack Watling and Nick Reynolds observed that "Interestingly, there is minimal interest among Russian crews in synchronising these effects with other activities or with deconflicting their effects" [63]. This lack of synchronisation may have followed a similar logic in the use of offensive cyberspace activities. The objective was to deploy malware extensively without attempting to achieve deep synchronisation across different domains. Such a course of action aligns with the principles of Russian warfare, which place importance on the initial phase of war, preemptive measures [64], and information operations conducted in the grey zone.

Conclusions

The shortcomings of the Russian army during the so-called Special Operation against Ukraine launched in February 2022 can be observed with the naked eye. However, in the cyber domain, there was one exception, indirectly indicating Russia's high offensive capabilities. Expert attention focused on the sole officially confirmed and successful offensive cyberattack on Viasat, a satellite internet provider. The objective of this attack was to undermine the Ukrainian military's command and control system (C2). Notably, this attack occurred just hours before the invasion commenced, garnering interest from Western analysts as an example of cross-domain coordination. While the internet blockade posed difficulties in defending Kiev during the early days of the war, it did not grant Russia enough of a military advantage to capture the Ukrainian capital or significantly influence the course of the conflict. The absence of other documented instances of effective Russian cyber operations during this conflict makes it easier to interpret Russian failures in cyberspace as part of the overall *bardak* within Russia. However, it appears that Russian strategic goals in cyberspace were much more modest than what Western experts had imagined.

The text argues that this is because the Russians acted in accordance with their strategic culture, wherein information warfare is crucial for hybrid warfare, but not instrumental in gaining territory. Offensive actions in cyberspace may hold tactical significance but lack strategic importance. The concept of cyberwar, as envisioned by Western analysts, involving offensive actions against the enemy's critical infrastructure during kinetic warfare, did not materialise. This was evident not only in 2022 but also in earlier conflicts such as the 2008 war with Georgia and the 2014 armed conflict with Ukraine. Russian offensive activities in cyberspace aimed at achieving strategic victory primarily involved mass malware attacks in the initial phase, but later

shifted towards intelligence activities and disinformation campaigns. Decisive cyberattacks are not the most important element of this strategy. It seems that Russia acknowledges the limited role of cyberspace in kinetic warfare, primarily focusing on intelligence and subversion, assigning more significance to it. And it will most likely stay that way in the future.

Acknowledgements

The publication has been supported by a grant from the Faculty of International and Political Studies under the Strategic Programme Excellence Initiative at Jagiellonian University.

References

[1] S. Gordon, E. Rosenbach, "America's Cyber-Reckoning," Foreign Affairs, Dec. 14, 2021. [Online]. Available: https://www.foreignaffairs.com/articles/unitedstates/2021-12-14/americas-cyber-reckoning [Accessed: May 01, 2023].

[2] N. Kostyuk, "Why Cyber Dogs Have Yet to Bark Loudly in Russia's Invasion of Ukraine," Texas National Security Review, Jun. 23, 2022. [Online]. Available: https://tnsr.org/2022/06/why-cyber-dogs-have-yet-to-bark-loudly-in-russias-invasionof-ukraine/ [Accessed: May 01, 2023].

[3] J. Lewis, "Cyber War and Ukraine," Center for Strategic and International Studies, Jun. 16, 2022. [Online]. Available: https://www.csis.org/analysis/cyber-war-and-ukraine [Accessed: May 01, 2023].

[4] M. Willett, "The Cyber Dimension of the Russia – Ukraine War," IISS, Oct. 06, 2022. [Online]. Available: https://www.iiss.org/blogs/survival-blog/2022/10/the-cyber-dimension-of-the-russia-ukraine-war [Accessed May 01, 2023].

[5] R. C. Maness, B. Valeriano, K. Hedgecock, J. Macias, B. Jensen, "Expanding the Dyadic Cyber Incident and Campaign Dataset (DCID): Cyber Conflict from 2000," The Cyber Defense Review, Sep. 22, 2022. [Online]. Available: https://cyberdefensereview.army.mil/CDR-Content/Articles/Article-View/Article/3500241/expanding-the-dyadic-cyber-incident-and-campaign-dataset-dcid-cyber-conflict-fr/ [Accessed Nov. 03, 2023].

[6] M. Hasian, S. T. Lawson, M. McFarlane, *The Rhetorical Invention of America's National Security State*. Lanham, MA: Lexington Books, 2015.

[7] K. Alexander, "Keith B. Alexander, Commander, U.S. Cyber Command, Memorandum for Record, Subject: United States Cyber Command (USCYBERCOM)

Commander's Strategic Assessment for Operating in Cyberspace – Preventing a Pearl Harbor Environment. Unclassified. National Security Archive," nsarchive.gwu.edu, Mar. 23, 2012. [Online]. Available: https://nsarchive.gwu.edu/document/21531-document-2-7. [Accessed: Nov. 03, 2023].

[8] S. Lawson, M. K. Middleton, "Cyber Pearl Harbor: Analogy, fear, and the framing of cyber security threats in the United States, 1991–2016," *First Monday*, vol. 24, no. 3, 2019, doi: 10.5210/fm.v24i3.9623.

[9] M. Libicki, "Defending Cyberspace and Other Metaphors," [Online]. Available: https://apps.dtic.mil/dtic/tr/fulltext/u2/a368431.pdf. [Accessed: Nov. 03, 2023].

[10] Army Cyber Institute, "What the Heck Is Threatcasting?," Sep. 15, 2017. [Online]. Available: https://cyber.army.mil/Library/Media-Coverage/Article/1342180/whatthe-heck-is-threatcasting/. [Accessed: May 01, 2023].

[11] J. Marks, "Here's what cyber pros are watching in the Ukraine conflict," Washington Post, Feb. 24, 2022. [Online]. Available: https://www.washingtonpost.com/politics/2022/02/24/heres-what-cyber-pros-are-watching-ukraine-conflict/. [Accessed: Aug. 02, 2023].

[12] W. Courtney, P. Wilson, "Expect 'shock and awe' if Russia invades Ukraine," The Hill, Dec. 08, 2021. https://thehill.com/opinion/international/584805-expect-shock-and-awe-if-russia-invades-ukraine/ [Accessed: Aug. 03, 2021].

[13] K. Giles, "Putin does not need to invade Ukraine to get his way," Chatham House – International Affairs Think Tank, Dec. 21, 2021. [Online]. Available: https://www.chathamhouse.org/2021/12/putin-does-not-need-invade-ukraine-get-his-way. [Accessed: May 01, 2023].

[14] D. Cattler, D. Black, "The Myth of the Missing Cyberwar," Foreign Affairs, Apr. 06, 2022. [Online]. Available: https://www.foreignaffairs.com/articles/ukraine/2022-04-06/myth-missing-cyberwar?check_logged_in=1#author-info [Accessed: May 01, 2023].

[15] J. Bateman, "Russia's Wartime Cyber Operations in Ukraine: Military Impacts, Influences, and Implications," Carnegie Endowment for International Peace, Dec. 16, 2022. [Online]. Available: https://carnegieendowment.org/2022/12/16/russia-s-wartime-cyber-operations-in-ukraine-military-impacts-influences-and-implications-pub-88657 [Accessed: Nov. 15, 2023].

[16] M. Miller, "Russia's cyberattacks aim to 'terrorize' Ukrainians," Politico, Jan. 11, 2023. [Online]. Available: https://www.politico.com/news/2023/01/11/russias-cyberattacks-aim-to-terrorize-ukrainians-00077561 [Accessed: Dec. 02, 2023].

[17] B. Smith, "Defending Ukraine: Early Lessons from the Cyber War," Jun. 22, 2022. [Online]. Available: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE50KOK. [Accessed: Nov. 07, 2023].

[18] J. Marks, "11 reasons we haven't seen big Russian cyberattacks yet," Washington Post, Mar. 03, 2022. [Online]. Available: https://www.washingtonpost.com/politics/2022/03/03/11-reasons-we-havent-seen-big-russian-cyberattacksyet/. [Accessed: Jun. 03, 2023].

[19] J. Bateman, N. Beecroft, G. Wilde, "What the Russian Invasion Reveals About the Future of Cyber Warfare," Carnegie Endowment for International Peace, Dec. 19, 2022. [Online]. Available: https://carnegieendowment.org/2022/12/19/whatrussian-invasion-reveals-about-future-of-cyber-warfare-pub-88667. [Accessed: Jul. 23, 2023].

[20] D. Dziwisz, "Cyber Pearl Harbor is Not Coming: US Politics Between War and Peace," *Politeja*, vol. 19, no. 4 (79), 2022, doi: 10.12797/politeja.19.2022.79.07.

[21] L. Maschmeyer, "A new and better quiet option? Strategies of subversion and cyber conflict," *Journal of Strategic Studies*, vol. 29, no. 1, pp. 1–25, 2022, doi: 10.1080/01402390.2022.2104253.

[22] S. Zilincik, I. Duyvesteyn, "Strategic studies and cyber warfare," Journal of Strategic Studies, vol. 11, pp. 1-22, 2023, doi: 10.1080/01402390.2023.2174106.

[23] L. Morris, M. Mazarr, J. Hornung, S. Pezard, A. Binnendijk, M. Kepe, "Gaining Competitive Advantage in the Gray Zone Response Options for Coercive Aggression Below the Threshold of Major War," RAND Corporation. [Online]. Available: https://www.rand.org/content/dam/rand/pubs/research_reports/RR2900/RR2942/RAND_RR2942.pdf. [Accessed: Nov. 03, 2023].

[24] G. Popp, S. Canna, "The Characterization and Conditions of the Gray Zone: A Virtual Think Tank Analysis (ViTTa) Prepared for Strategic Multi-Layer Assessment Gray Zone Conflicts, Challenges, and Opportunities: A Multi-Agency Deep Dive Assessment," NSI, Inc. [Online]. Available: https://nsiteam.com/social/wp-content/uploads/2017/01/Final_NSI-ViTTa-Analysis_The-Characterization-and-Conditionsof-the-Gray-Zone.pdf. [Accessed: Nov. 03, 2023].

[25] Office of the Secretary of Defense, "Office of the Secretary of Defense Annual Report to Congress: Military and Security Developments Involving the People's Republic of China," 2020. [Online]. Available: https://media.defense.gov/2020/Sep/01/2002488689/-1/-1/1/2020-DOD-CHINA-MILITARY-POWER-REPORT-FINAL.PDF. [Accessed: Nov. 03, 2023].

[26] S. Takashi, "Overview of current research," Nakasone Peace Institute, 2020. [Online]. Available: https://www.npi.or.jp/en/research/NPI_Research_Note_20201005.pdf. [Accessed: Nov. 03, 2023].

[27] J. R. Van der Velde, "Make Cyberspace Great Again Too!," Real Clear Defense, Jul. 23, 2018. [Online]. Available: https://www.realcleardefense.com/articles/2018/07/23/make_cyberspace_great_again_too_113634.html. [Accessed: Jul. 23, 2023].

[28] G. Casey, "Aug. 14, 2007 – Remarks at the National Press Club," [Online]. Available: https://www.army.mil/article/4436/aug_14_2007_remarks_at_the_national_press_club. [Accessed: Jul. 15, 2023].

[29] J. Darczewska, "The devil is in the details. Information warfare in the light of Russia's military doctrine," Warsaw: Centre for Eastern Studies, May 2015. [Online]. Available: https://www.osw.waw.pl/sites/default/files/pw_50_ang_thedevil-is-in_net.pdf. [Accessed: Nov. 03, 2023].

[30] D. Minic, "How the Russian army changed its concept of war, 1993–2022," *Journal of Strategic Studies*, pp. 1–35, 2023, doi: 10.1080/01402390.2023.2199445.

[31] "Doctrine of Information Security of the Russian Federation". (Dec. 05, 2016). http://www.scrf.gov.ru/security/information/DIB_engl/ [Accessed: Nov. 02, 2023].

[32] "Strategy of National Security of the Russian Federation". (2021). [Online]. Available: https://paulofilho.net.br/wp-content/uploads/2021/10/National_Security_Strategy_of_the_Russia.pdf [Accessed: Oct. 10, 2023].

[33] M. Skak, "Russian strategic culture: the role of today's *chekisty*," *Contemporary Politics*, vol. 22, no. 3, pp. 324–341, 2016, doi: 0.1080/13569775.2016.1201317.

[34] C. Pursiainen, "Russia's Critical Infrastructure Policy: What do we Know About it?," *European Journal for Security Research*, vol. 6, 2020, doi: 10.1007/s41125-020-00070-0.

[35] K. B. Payne, J. S. Foster, "Russian Strategy Expansion, Crisis and Conflict," Comparative Strategy, vol. 36, no. 1, pp. 1–89, 2017, doi: 10.1080/01495933.2017.1277121.

[36] S. Fabian, "The Russian hybrid warfare strategy – neither Russian nor strategy," *Defense & Security Analysis*, vol. 35, no. 3, pp. 308–325, 2019, doi: 10.1080/14751798.2019.1640424.

[37] J. A. Kerr, "Concept Misalignment and Cyberspace Instability: Lessons from Cyber-Enabled Disinformation," in *Cyberspace and Instability*, R. Chesney, J. Shires, M. Smeets, Eds: Edinburgh University Press, 2023, pp. 99–126.

[38] M. Kofman, A. Fink, D. Gorenburg, M. Chesnut, J. Edmonds, J. Waller, "Russian Military Strategy: Core Tenets and Operational Concepts," [Online]. Available: https://www.cna.org/archive/CNA_Files/pdf/russian-military-strategy-core-tenets-and-operational-concepts.pdf. [Accessed: May 01, 2023].

[39] R. Thornton, M. Miron, "Winning Future Wars: Russian Offensive Cyber and Its Vital Importance," [Online]. Available: https://cyberdefensereview.army.mil/Portals/6/Documents/2022_summer_cdr/09_Thorton_Miron_CDR_V7N3_Summer_2022.pdf?ver=0LhzDv4-cUkzkAqiTz401g%3D%3D. [Accessed: May 01, 2023].

[40] J. Hakala, J. Melnychuk, "Russia's Strategy in Cyberspace," NATO Strategic Communications Centre of Excellence, 2021. [Online]. Available: https://stratcomcoe.org/cuploads/pfiles/Nato-Cyber-Report_11-06-2021-4f4ce.pdf. [Accessed: Jul. 12, 2023].

[41] K. Giles, A. Seaboyer, "The Russian Information Warfare Construct," Defence Research and Development Canada, 2019. [Online]. Available: https://cradpdf.drdc-rddc.gc.ca/PDFS/unc341/p811007_A1b.pdf. [Accessed: Oct. 10, 2023].

[42] K. Giles, "Russia's 'New' Tools for Confronting the West Continuity and Innovation in Moscow's Exercise of Power Russia's 'New' Tools for Confronting the West," [Online]. Available: https://www.chathamhouse.org/sites/default/files/publications/2016-03-russia-new-tools-giles.pdf [Accessed: Sep. 09, 2023].

[43] C. Reach, A. A. Blanc, E. Geist, "Russian Military Strategy Organizing Operations for the Initial Period of War Research Report," Santa Monica: RAND Corporation, 2022. [Online]. Available: https://www.rand.org/content/dam/rand/pubs/research_reports/RRA1200/RRA1233-1/RAND_RRA1233-1.pdf. [Accessed: Jun. 12, 2023].

[44] C. Reach, A. Demus, M. Grisé, K. Holynska, C. Lynch, D. Massicot, D. Woodworth, "Russia's Evolution Toward a Unified Strategic Operation. The Influence of Geography and Conventional Capacity Research Report," RAND Corporation, Santa Monica, 2023. [Online]. Available: https://apps.dtic.mil/sti/trecms/pdf/AD1193347.pdf. [Accessed: Nov. 03, 2023].

[45] M. J. Kari, K. Pynnöniemi, "Theory of strategic culture: An analytical framework for Russian cyber threat perception," *Journal of Strategic Studies*, vol. 46, no. 1, pp. 56–84, 2019, doi: 10.1080/01402390.2019.1663411.

[46] S. P. White, "Understanding Cyberwarfare: Lessons from the Russia-Georgia War," Modern War Institute, Mar. 20, 2018. [Online]. Available: https://mwi.westpoint.edu/understanding-cyberwarfare-lessons-russia-georgia-war/. [Accessed: Oct. 06, 2023].

[47] E. Gartzke, "The Myth of Cyberwar: Bringing War in Cyberspace Back Down to Earth," *International Security*, vol. 38, no. 2, pp. 41–73, 2013, doi: 10.1162/ISEC_a_00136.

[48] K. Giles, P. Hanson, R. Lyne, J. Nixey, J. Sherr, A. Wood, "Chatham House Report: The Russian Challenge," London: Chatham House, 2015. [Online]. Available: https://www.chathamhouse.org/sites/default/files/field/field_document/20150605RussianChallengeGilesHansonLyneNixeySherrWoodUpdate.pdf. [Accessed: Oct. 10, 2023].

[49] P. Tucker, "Why Ukraine Has Already Lost the Cyberwar, Too," *Defense One*, Apr. 28, 2014. [Online]. Available: https://www.defenseone.com/technology/2014/04/whyukraine-has-already-lost-cyberwar-too/83350/ [Accessed: Oct. 03, 2023].

[50] J. A. Lewis, "Compelling Opponents to Our Will': the Role of Cyber Warfare in Ukraine," in *Cyber War in Perspective: Russian Aggression against Ukraine*, K. Geers, Ed. Tallinn: NATO Cooperative Cyber Defence Centre of Excellence, 2015, pp. 39–47.

[51] J. J. Driedger, "Russian Active Measures against Ukraine (2004) and Estonia (2007)," in Russian Active measures – yesterday, today, tomorrow, O. Bertelsen, Ed. Stuttgart: Columbia University Press, 2021, pp. 177-213.

[52] G. B. Mueller, B. Jensen, B. Valeriano, R. C. Maness, J. M. Macias, "Cyber Operations during the Russo-Ukrainian War," Center for Strategic and International Studies, Jul. 13, 2023. [Online]. Available: https://www.csis.org/analysis/cyber-operations-during-russo-ukrainian-war. [Accessed: Nov. 03, 2023].

[53] J. J. Driedger, "Risk acceptance and offensive war: The case of Russia under the Putin regime," *Contemporary Security Policy*, vol. 44, no. 2, pp. 199–225, 2023, doi: 10.1080/13523260.2023.2164974.

[54] T. Bukkvoll, "Why Putin Went to War: Ideology, Interests and Decision-making in the Russian Use of Force in Crimea and Donbas," Contemporary Politics, vol. 22, no. 3, pp. 267-282, 2016, doi: 10.1080/13569775.2016.1201310.

[55] K. Geers, "Computer Hacks in the Russia-Ukraine War," DEFCON Conference, Aug. 11, 2022. [Online]. Available: https://media.defcon.org/DEF%20CON%2030/DEF%20CON%2030%20presentations/Kenneth%20Geers%20-%20Computer%20Hacks%20in%20the%20Russia-Ukraine%20War%20-%20paper.pdf [Accessed: Oct. 10, 2023].

[56] "An Overview of Russia's Cyberattack Activity in Ukraine Special Report: Ukraine Digital Security Unit", Microsoft Digital Security Unit, Apr. 27, 2022. [Online]. Available: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4Vwwd [Accessed: Oct. 10, 2023].

[57] L. Maschmeyer, "The Subversive Trilemma: Why Cyber Operations Fall Short of Expectations," *International Security*, vol. 46, no. 2, pp. 51–90, 2021, doi: 10.1162/isec_a_00418.

[58] H. Lin, "Russian Cyber Operations in the Invasion of Ukraine," The Cyber Defense Review, vol. 7, no. 4, pp. 31-46, 2022. [Online]. Available: https://www.jstor.org/stable/48703290. [Accessed: Jul. 07, 2023].

[59] S. J. Freedberg Jr., "Cyber lessons from Ukraine: Prepare for prolonged conflict, not a knockout blow," Breaking Defense, May 01, 2023. [Online]. Available: https://breakingdefense.com/2023/05/cyber-lessons-from-ukraine-prepare-for-prolonged-conflict-not-a-knockout-blow/. [Accessed: Jul. 07, 2023].

[60] T. Starks, A. Schaffer, "Did Russia Mess up Its Cyberwar with Ukraine before It Even Invaded?," Washington Post, Sep. 04, 2022. [Online]. Available: https://www.washingtonpost.com/politics/2022/08/04/did-russia-mess-up-its-cyberwar-with-ukraine-before-it-even-invaded/. [Accessed: Sep. 10, 2023].

[61] A. Levite, "Integrating Cyber into Warfighting: Some Early Takeaways from the Ukraine Conflict," Carnegie Endowment for International Peace, Apr. 18, 2023. https://carnegieendowment.org/2023/04/18/integrating-cyber-into-warfighting-some-early-takeaways-from-ukraine-conflict-pub-89544 [Accessed: Jul. 07, 2023].

[62] D. E. Sanger, J. E. Barnes, "Many Russian Cyberattacks Failed in First Months of Ukraine War, Study Says," The New York Times, Jun. 22, 2022. [Online]. Available: https://www.nytimes.com/2022/06/22/us/politics/russia-ukraine-cyberattacks.html. [Accessed: Jun. 10, 2023].

[63] J. Watling, N. Reynolds, "Meatgrinder: Russian Tactics in the Second Year of Its Invasion of Ukraine," Royal United Services Institute for Defence and Security Studies, May 19, 2023. [Online]. Available: https://www.rusi.org/explore-our-research/publications/special-resources/meatgrinder-russian-tactics-second-year-its-invasion-ukraine. [Accessed: Oct. 10, 2023]

[64] T. Thomas, "McLean, VA Russian Military Thought: Concepts and Elements," MITRE CORP MCLEAN VA, 2019. [Online]. Available: https://www.mitre.org/sites/default/files/2021-11/prs-19-1004-russian-military-thought-concepts-elements.pdf. [Accessed: Oct. 11, 2023].

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Tell Me Where You Live and I Will Tell Your P@Ssw0rd: Understanding the Macrosocial Variables Influencing Password's Strength

Andreanne Bergeron | GoSecure; University of Montreal, Canada,
ORCID: 0000-0001-9013-6662

Abstract

Users' habits in relation to cybersecurity are frequently examined from the micro perspective, using survey results to obtain impactful variables from individuals, focusing on usability and security factors of passwords. In this paper, the influence of macrosocial factors on password strength is studied in order to offer a global comprehension of the influence of the environment on users. Using the list of the 200 most common passwords by countries released by NordPass in 2021, logistic regression has been used to predict macrosocial variable influencing password strength. Results show that (1) Literacy level of a population; (2) Voice and accountability; (3) Level of global cybersecurity; and (4) Level of data breaches exposure significantly predict users' password strength performance. The author discusses the impact of government on password hygiene of users hoping to influence the development of policies around cyber security configurations and investment set by nations and institutions.

Keywords
password, macrosocial influence, authentication, users' behaviour, users protection

Received: 28.07.2023

Accepted: 06.11.2023

Published: 27.11.2023

Cite this article as:
A. Bergeron "Tell Me Where You Live and I Will Tell Your P@Ssw0rd: Understanding the Macrosocial Variables Influencing Password's Strength," ACIG, vol. 2, no. 1, 2023, DOI: 10.60097/ACIG/162863

Corresponding author:
Andreanne Bergeron, GoSecure; University of Montreal, Canada; ORCID: 0000-0001-9013-6662; E-MAIL: andreanne.bergeron.5@umontreal.ca

1. Introduction

Passwords are words, strings of characters, or some form of interactive message used to prove identity and gain access to a resource or a place. They constitute the first line of defence for computer-based technologies and were used for millennia as the Roman military were reportedly using passwords to distinguish allies from enemies [1]. Today, even if attack-resistant validation schemes exist, passwords constitute the most popular strategy of authentication.

People usually have a multitude of different passwords and when they create them, they often use a strategy to make it easy to remember [2, 3]. Past studies have shown that users intend to choose weak passwords, which are usually easy to be remembered but vulnerable to be guessed [4, 5]. Also, study reveals that textual passwords are often reused, which have been shown to be an important security threat of passwords [6].

Researchers demonstrated the influence of a person's environment and exposure to the Internet on their online security behaviour [7, 8]. Password creation strategy, defined as the active approaches that can be used by a password creator to create memorable passwords [9, 3], also seems to be influenced by a person's environment. For example, it was identified that students from the United States have a higher risk perception toward surveillance than students from the United Kingdom [8]. Also, Yang et al. [10] discuss the cultural influence in password choice. They explain the weak passwords strength level of Chinese by the rapid growth of Internet users and e-commerce markets in China. They hypothesize that providers may not have paid enough attention to security issues because of the focus on market expansion. The results of the aforementioned studies suggest that there is a structural difference in cybersecurity habits between countries [11].

The present study aims to explore the various macrosocial elements contributing to the structural difference between countries in users' choice of password. The contribution of governments to the problem or to the solutions can be evaluated through this assessment. In order to observe countries' differences, the password strength performance of users will be compared to macrosocial variables that could influence password creation strategy.

1.1. Macrosocial Variables Influencings Users

Few studies show evidence that there is a structural difference between countries in password habits (e.g., [11]) but

the macrosocial variables influencing it have rarely been directly tested. To explore the different variables that could play a role in password habits, the literature on macrosocial variables influencing the different aspect of technology, like the use of Internet in general, is considered. Several macrosocial elements might be taken into consideration when evaluating the reasons why users have different levels of performance according to their environment. First, if there is a difference in cybersecurity habits between countries, the characteristics of the government might be an element influencing users. Second, the characteristics of the population, which is directly related to users, would also be an element explaining the impact of the environment. Finally, external variables like cyber-attacks and the level of cyberattack victimization of a country might also be a part of the explanation.

Characteristics of the government. The economic aspect of a government might influence Internet habits. Prior studies have found that a country's economic development level helps predict the use of internet in a society [12]. Gross domestic product (GDP) is the most closely watched and important economic indicator and considers different variables about a country's economy, including its consumption and investment [13, 14]. It could be hypothesized that economic indicator would influence not only the use of internet, but other habits related to them.

Along with the economy, the investment and the commitment of countries to cybersecurity is an important variable to consider as the relation is more direct. Researchers have found that when a country invests and commit into the cybersecurity sector, the annual losses due to cybercrime over the country's Gross National Income decreases [15]. The investments in cybersecurity can include education and tools to help users to more efficiently manage their Internet use.

Characteristics of the population. Digital skills and the overall ability to use the internet are two elements that are directly linked to literacy. Internet users are reading expository text in a hypertext format where ideas are connected by links, headings, icons, and graphics; those elements necessitate similar reading strategies as those used with print text reading [16]. In other words, to seek, evaluate, and use information found on the Internet, readers must navigate through Internet text and apply their knowledge of the reading process. To understand correctly what a password is and to write one, people need to read. Research has shown that password security practices typically conflict with general usability principles [17]. The challenges faced by low-literacy users when creating and managing passwords

are likely to extend beyond those experienced by the public. Literacy level affects password habits [18].

External events: Data breaches. According to the Identity Theft Resource Centre's Annual Data, there were 1862 data breaches in 2021. Researchers have shown that the United States was highly represented in data breaches, and they explain this by their high level of economic activity as well as by their relatively high notification rates they have compared to other countries [19]. Luxemburg, Canada and Great Britain follow the United States in the list of countries most affected by data breaches [19]. When calculating the country-based probability variable, another study shown that France and Brazil have relatively higher probability of data breaches than the other countries [20]. Researchers state that the probability of a data breach is influenced by the country in which it happens [20].

There seems to be a relationship between the influence that data breaches might have on users and their habits. Campbell et.al. [21], examined the stock market reaction to newspaper reports of information security breaches at 38 publicly traded U.S. corporations during the period January 1, 1995 to December 31, 2000. Among the 43 different events, the authors found a highly significant negative market reaction to information security breaches involving unauthorized access to confidential data. Moreover, in their study on 6,000 users from the United States, after a data breach notification, victims changed their password or PIN (51%) or switched to a new account (24%) [22]. The literature suggests that users actively assess the consequences of breaches and react accordingly.

1.2. Aim of the study

Users' habits in relation to cybersecurity is frequently examined from the micro perspective, using survey results to obtain impactful variable from individuals, focusing on usability and security factors of passwords [23, 24]. In this paper, the influence of macrosocial elements on password strength[1] is studied in order to offer a global comprehension of the influence of the environment on users. Exploring the different concept of technology and their flaws at the country level encourages future development of new technologies and improve related capital investments (e.g., [28, 29]).

A descriptive analysis of leaked lists of passwords in 2021 is conducted to determine which macrosocial variables would be included in the model and therefore play a formative role in how users formulate their passwords across countries. Then, a prediction model help

1 — Strong passwords are usually characterized by larger number of characters, containing upper and lowercase letters, numbers and special characters [25]. Also, a strong password should avoid using dictionary words [26, 27].

identify the extent to which variables influence password strength. This study is innovative as it allows to investigate trends in password formulation with regard to social context. The impact of our study is a move toward a better understanding of human behaviour in the context of password formulation specifically, to enable the future crafting of more targeted cybersecurity interventions that would lead to positive online behavioural change.

2. Method

2.1 Sample

Each year, the company NordPass release a list of the 200 most common passwords by country. The list of passwords is compiled using the many cybersecurity incidents (data breaches containing users' password) that occur in 2021. In total, the list rose from 4 terabytes of information and contain 49 countries. The complete list of countries can be found in Appendix A.

The list comprises between 169,656 and 146,837,497 users' account per country. The average time to crack passwords is 2082684.368 seconds (range from 0 to 3,214,080,000 seconds). The majority of passwords included in the list can be cracked in less than a minute (61%). The fact that the mean time to crack a password is high in a country means that high quality passwords were included in the 200 most commonly used: the password can be common, but the overall strength is high.

2.2. Measures

In order to account for the strength of passwords, the mean time to crack the password, which was already included in NordPass passwords list, was observed. Then, several macrosocial variables have been considered to create a model explaining the level of password strength. A total of 29 different measures have been scrutinized in the exploration of possible model explaining performance of countries in password strength. In order to maintain a low risk of overfitting in the model, a limited number of variables can be inserted in relation to the number of cases (49 countries). The literature reports that one predictive variable can be studied for every ten events (i.e., number of countries) [30, 31]. The complete list of measures that have been considered can be found in Appendix B. In order to determine the five variables to be entered in the model, the first step was to do a correlation matrix. This allowed to avoid highly correlated variables to be entered the model together. Then,

different models were tested using an amalgam of variables from the list with a special attention to the important aspect identified in the literature review. The contribution of the variable to the model were very stable and most of them have been chosen because they were predicting password strength. The five variables chosen to enter the final model are named and defined below.

Voice & Accountability (2020)

It is one of six components of governance indicator as stipulated by the World Bank. It reflects perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media.

Global Cybersecurity Index (2020)

The Global Cybersecurity Index (GCI) is a trusted reference that measures the commitment of countries to cybersecurity at a global level and is composed of 25 indicators that monitor and compare the level of the cybersecurity commitment of countries with regard to the five pillars – (i) Legal Measures, (ii) Technical Measures, (iii) Organizational Measures, (iv) Capacity Development, and (v) Cooperation – and then aggregated into an overall score. It represents the most comprehensive measures of cybersecurity commitment of countries compared to many other measures that are published by corporations [15].

Cybersecurity Exposure Index (2020)

The Cyber Exposure Index is based on data collected from publicly available sources in the dark web and deep web and from data breaches. From this data, signs of sensitive disclosures, exposed credentials and hacker-group activity against companies are identified.

Literacy (2022)

This measures the percentage of adults in a country who are able to read and write their common language. A higher literacy rate is an indication of higher standards of education and the good ability of the population to find formal employment.

GDP per Capita (2020)

Gross domestic product (GDP) is the standard measure of the value added created through the production of goods and services in a country during a certain period. As such, it also measures the income earned from that production, or the total amount spent on final goods and services (fewer imports).

2.3. Analysis

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple regression is an extension of linear regression that uses just one explanatory variable. MLR assumes that there is a linear relationship between the dependent variable and the independent variables. It also assumes that the data should not show multicollinearity, which occurs when the independent variables (explanatory variables) are highly correlated. The amount of error in the residuals is similar at each point of the linear model, the observations should be independent of one another and occurs when residuals are normally distributed [32]. All those assumptions have been tested through data observation. The software IBM SPSS 28 was used to do the analysis.

3. Results

Multiple linear regression was used to test if the five macrosocial variables under study significantly predicted password strength. The overall regression was statistically significant ($R^2 = 0.36$, $F = 23.46$, $p = < 0.004$). The model is presented in Table 1.

Table 1. Multiple Linear Regression Results (standard deviation from the mean).

(Constant)	-84076485.358*** (23719091.430)
Voice and accountability	141343.544** (50531.714)
Global Cybersecurity Index	384493.225** (114063.016)
Cybersecurity Exposure Index	49756239.387*** (12339781.870)
Literacy	288067.744* (156148.862)
GDP per Capita	27.981 (71.114)
R-squared	0.36
Number of observations	49
***p<0.001; **p<0.05; *p<0.1	

It was found that Voice and accountability (β=141343.544, p=0.008), Global Cybersecurity Index (β=384493.225, p=0.002), Cybersecurity Exposure Index (β=49756239.387, p=0.000), and level of literacy (β=288067.744, p=0.072) significantly predicted password strength. It was also found that the GDP per capita (β=27.981, p=0.696) did not significantly predict password strength.

4. Discussion

The analysis of the present study help identifies a variety of different macrosocial measures significantly predicting password strength of users: literacy, voice and accountability, level of global cybersecurity, and the level of cybersecurity exposure. Considering past literature on the subject, one variable was surprisingly not associated with an increase password strength, that is, the GDP per capita. Each of those measures are presented in this section in the light of previous work through broader categories: Characteristics of the government, characteristics of users and external variables.

4.1. Characteristics of the Government

Freedom in a country has shown to have an impact on Internet use. Voice and accountability indicator reflects perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media. Researchers suggest that greater levels of Internet diffusion are associated with greater levels of voice and accountability [33, 34]. Musa et al. [35] argue that developing countries are more resistant than developed countries to the introduction of technologies that can be used to fight corruption, such as Internet-based technologies. Beyond the use of internet, the impact of freedom is seen on cybersecurity. A strong positive association has been shown between Cybersecurity Capacity Scale and cross-national indicators of citizen perceptions of having voice and accountability [36]. The result of the present study confirms the impact of voice and accountability on password performance as this variable is a good predictor of password strength.

The adoption of technology in a country has been proven to be impacted by many factors including its economic development and growth [37]. Compared to more developed countries, countries that are less developed possess inferior infrastructure, less effective manpower (partly because of low education levels), and business models that have not shifted from the industrial age to the information age [38]. The wealth disparity has also been noted to impact technology adoption, although previous studies have examined the wealth disparity from a micro level [39 – 42]. The result of the present study indicates that wealth disparity does not influence the strong password hygiene as GDP was not significantly predicting password strength. This result might be explained by the sector in which developed countries invest but also other influencing variable like experience in the IT sector. Past studies have shown that countries need to acquire experience with IT before investments begin to

reward the country economically [43]. Benefiting from resources is not enough to explain effective use of technology.

Even if the GDP is not a significant element in citizen password strength, the level of investment of a government in cybersecurity has an impact on security of users. The Global Cybersecurity Index (GCI) is a trusted reference that measures the commitment of countries to cybersecurity at a global level. Researchers have found that as the score for the Global cybersecurity increases, the annual losses due to cybercrime for each country over their Gross National Income decreases [15]. The literature shows that commitment of countries to fight against cybersecurity is profitable economically. The present study goes further by showing the impact on users by demonstrating that this type of investment predicts better password strength performance.

4.2. Characteristics of Users at a Macro-Level

Literacy is an important aspect to consider in this study as it is directly connected to the use of technologies. To seek, evaluate, and use information found on the Internet, readers must navigate through Internet text and apply their knowledge of the reading process [16]. Today's definition of literacy is being broadened to include "literacy skills necessary for individuals, groups, and societies to access the best information in the shortest time to identify and solve the most important problems and then communicate this information" [44]. Most knowledge of late trends on technology is acquired by information found on the Internet. Because bei;00ng knowledgeable is closely related to the capacity to acquire this knowledge (e.g., being able to read), people with low level of literacy can hardly adapt. The challenges faced by low-literacy users when creating and managing passwords are documented and research indicates that they are higher than the general population [18]. Research shows that when users' level of cyber security knowledge increases, so does their cybersecurity behaviour contributing to good hygiene [45]. However, if users are not able to get this information about cybersecurity because of their inability to read, their security will be impacted. The results of the present study are therefore not surprising: when the level of literacy of a population increases, the strength of passwords also increases.

4.3. External Variables Influencing Countries

The results show that the number of cybersecurity incidents exposure of a country is positively associated with password

change. The more a country is under attack, the more people use strong passwords. This suggests that people might be sensible to the importance of protecting data with strong passwords when they are exposed to more cybersecurity incidents. Users are well aware of the meaning of a data breach [46], and it influences their behaviour. For example, there is a highly significant negative market reaction to information security breaches involving unauthorized access to confidential data [21]. In their study on 6,000 users from the United States, after a data breach notification, victims changed their password or PIN (51%) or switched to a new account (24%) [22]. Users are also recognized to be comfortable with proactive password resetting in the event of reuse and sharing information with other identity providers [46]. Therefore, users are aware of what will protect them and are more likely to do it when they are increasingly exposed to incidents. This demonstrates the resilience of users when they live in hostile environment but also the importance of making this information public as this knowledge is a protective factor for users. Mandatory reporting of data breaches introduced in Canada in 2018 [47] might be contributing solution to protect users.

4.4. Limitations

The set of data taken from Nordpass present important limitations as the method used to estimate the time to crack is unspecified. The list was investigated because the mean time to crack appears to be high. Some passwords from the list were weak (e.g., kallynlavallee) but were associated with a cracking time of more than 100 years. This is considered an important limitation of the dataset. However, the unspecified method is used consistently across the countries. Therefore, the metric could be used for the comparative analysis as it is consistent and can be relied upon.

Also, this study takes into consideration a macro perspective of the password strength, but a myriad of element can influence users' choices. The objective of the present study was to explore the influence of different large-scale policies and not individuals' decisional process.

5. Conclusion

The present study helps understand the importance of macrosocial variables on predicting password strength of users. It points toward the fact that some characteristics of the government influences password strength performance of users. For example,

democratic countries and countries in which the government invests in cybersecurity increase the password performance of users. The economic commitment of countries to fight against cybersecurity has been proven to be profitable economically and this study show that it is also associated with password strength of their citizens. Government has an important role to play on the cyber-protection of users whether it is direct (by investing in cyber security) or indirect (by prioritizing democracy and education).

Another important element raised by the present study is that exposure to data breaches increases the strength of user's password. This can be explained by the fact that the population adapt to the threat and this behaviour point toward the importance of mandatory reporting of data breaches by organizations. If they are confronted to mandatory reporting, users are more likely to know about the breaches and continue to adapt their behaviour and it becomes a protection factor.

Through a better understanding of human behaviour in the context of password formulation, our research focuses on identifying common denominators in behaviour that can lead to increased user vulnerabilities in online password formulation. The novelty of our exploratory research lies in our attempt to understand macrosocial variable associated with cybersecurity. The implication of this study concerns the development of policies around cyber security configurations and investment set by nations and institutions.

Funding

This work did not benefit from any funding.

Disclosure statement

The author has no conflict of interest.

References

[1] E. E. Best, "The literate Roman soldier," *The Classical Journal*, vol. 62, no 3, pp. 122–127, 1966.

[2] E. Stobert and R. Biddle, "The password life cycle: user behaviour in managing passwords," presented at the 10th Symposium on Usable Privacy and Security (SOUPS 2014), 2014. [Online]. Available: https://www.usenix.org/conference/soups2014/proceedings/presentation/stobert. [Accessed: July 28, 2023].

[3] B. Ur, S. M. Segreti, L. Bauer, N. Christin, L. F. Cranor, S. Komanduri, D. Kurilova, M. L. Mazurek, W. Melicher, R. Shay, "Measuring Real-World Accuracies and Biases in Modeling Password Guessability," presented at the 24th USENIX Security Symposium, 2015. [Online]. Available: https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/ur. [Accessed: July 28, 2023].

[4] A. Das, J. Bonneau, M. Caesar, N. Borisov, X. Wang, "The tangled web of password Reuse," in Proc. NDSS, 2014. [Online]. Available: https://www.cs.umd.edu/class/spring2017/cmsc818O/papers/tangled-web.pdf. [Accessed: July 28, 2023].

[5] J. Yan, A.F. Blackwell, R.J. Anderson, A. Grant, "Password memorability and security: empirical results," IEEE Security & Privacy, vol. 2, no. 5, pp. 25–31, 2004. [Online]. Available: doi: 10.1109/MSP.2004.81.

[6] W. Han, Z. Li, M. Ni, G. Gu, W. Xu, "Shadow attacks based on password reuses: A quantitative empirical view," IEEE Transactions on Dependable and Secure Computing, vol. 15, no. 2, pp. 309–320, 2018. [Online]. Available: doi: 10.1109/TDSC.2016.2568187.

[7] L. Bosnjak, B. Brumen, "What do students do with their assigned default passwords?," in 39th International Convention on Information and Communication Technology, Electronics and Microelectronics, 2016, pp. 1430–1435.

[8] P. Van Schaik, D. Jeske, J. Onibokun, L. Coventry, J. Jansen, P. Kusev, "Risk perceptions of cyber-security and precautionary behaviour," *Computer and Human Behavior*, vol. 75, pp. 547–559, 2017, doi: 10.1016/j.chb.2017.05.038.

[9] M. Zviran, W.J. Haga, "Cognitive passwords: The key to easy access control," *Computers & Security*, vol. 9, no. 8, pp. 723–736, 1990, doi: 10.1109/JCIT.1990.128279.

[10] C. Yang, J. L. Hung, Z. Lin, "An analysis view on password patterns of Chinese internet users," *Nankai Business Review International*, 2013, doi: 10.1108/20408741311303887.

[11] V. Nedvěd, "Careless society: Drivers of (un) secure passwords," M.A. thesis, Charles University, Prague, 2021. [Online]. Available: https://dspace.cuni.cz/handle/20.500.11956/126879. [Accessed: July 28, 2023].

[12] J. Corrales, F. Westhoff, "Information technology adoption and political regimes," *International Studies Quarterly*, vol. 50, no. 4, pp. 911–933, 2006, doi: 10.1111/j.1468-2478.2006.00431.x.

[13] M. Kummu, M. Taka, J.H. Guillaume, "Gridded global datasets for gross domestic product and Human Development Index over 1990–2015," *Scientific data*, vol. 5, no.1, pp. 1–15, 2018, doi: 10.1038/sdata.2018.4.

[14] L. Fioramonti, L. Coscieme, L.F. Mortensen, "From gross domestic product to wellbeing: How alternative indicators can help connect the new economy with the Sustainable Development Goals," *The Anthropocene Review*, vol. 6, no. 3, pp. 207–222, 2019, doi: 10.1177/2053019619869947.

[15] K. Farahbod, C. Shayo, J. Varzandeh, "Cybersecurity indices and cybercrime annual loss and economic impacts," *Journal of Business and Behavioral Sciences*, vol. 32, no. 1, pp. 63–71, 2020.

[16] E. Schmar-Dobler, "Reading on the Internet: The link between literacy and technology," *Journal of adolescent & adult literacy*, vol. 47, no. 1, pp. 80–85, 2003.

[17] D. Weirich, M.A. Sasse, "Pretty good persuasion: a first step towards effective password security in the real world," *Proceedings of the 2001 workshop on New security paradigms*, 2001. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/508171.508195. [Accessed : July 28, 2023].

[18] C. Rinn, K. Summers, E. Rhodes, J. Virothaisakun, D. Chisnell, "Password creation strategies across high – and low-literacy web users," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp.1–9, 2016, doi: 10.1002/pra2.2015.145052010052.

[19] K.M. Hogan, G.T. Olson, M. Angelina. (2020*). A comprehensive analysis of cyber data breaches and their resulting effects on shareholder wealth.* [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3589701. [Accessed : July 28, 2023].

[20] A. M. Algarni, V. Thayananthan, Y. K. Malaiya, "Quantitative assessment of cybersecurity risks for mitigating data breaches in business systems," *Applied Sciences*, vol. 11, no. 8, pp. 3678, 2021, doi: 10.3390/app11083678.

[21] K. Campbell, L.A. Gordon, M.P. Loeb, L. Zhou, "The economic cost of publicly announced information security breaches: empirical evidence from the stock market," *Journal of Computer security*, vol. 11, no. 3, pp. 431–448, 2003, doi: 10.3233/Jcs-2003-11308.

[22] L. Ablon, P. Heaton, D.C. Lavery, S. Romanosky, *Consumer attitudes toward data breach notifications and loss of personal information*. Santa Monica: Rand Corporation, 2016.

[23] C. Braz, A. Seffah, D. M'Raihi, "Designing a trade-off between usability and security: a metrics based-model," IFIP Conference on human-computer interaction, Rio de Janeiro, 2007, pp.114–126. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-74800-7_9. [Accessed : July 28, 2023].

[24] N. Gunson, D. Marshall, H. Morton, M. Jack, "User perceptions of security and usability of single-factor and two-factor authentication in automated telephone banking," *Computer Security*, vol. 30, no. 4, pp. 208–220, 2011, doi: 10.1016/j.cose.2010.12.001.

[25] D. Florencio, C. Herley, "A large-scale study of web password habits," *Proceedings of the 16th international conference on World Wide Web,* 2007. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/1242572.1242661. [Accessed: July 28, 2023].

[26] A. K. Kyaw, F. Sioquim, J. Joseph, "Dictionary attack on Wordpress: Security and forensic analysis," Second International Conference on Information Security and Cyber Forensics (InfoSec), Cape Town, 2015, pp. 158–164. [Online]. Available: https://ieeexplore.ieee.org/document/7435522 [Accessed: July 28, 2023].

[27] A. Narayanan, V. Shmatikov, "Fast dictionary attacks on passwords using time-space tradeoff," Proceedings of the 12th ACM conference on Computer and communications security, Alexandria, 2005, pp. 364–372. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/1102120.1102168. [Accessed: July 28, 2023].

[28] A. P. H. de Gusmão, M. M. Silva, T. Poleto, L. C., e Silva, A. P. C. S. Costa, "Cybersecurity risk analysis model using fault tree analysis and fuzzy decision theory," *International Journal of Information Management*, vol. 43, pp. 248–260, 2018, doi: 10.1016/j.ijinfomgt.2018.08.008.

[29] H. Taherdoost, "A review of technology acceptance and adoption models and Theories," *Procedia manufacturing*, vol. 22, pp. 960–967, 2018, doi: 10.1016/j.promfg.2018.03.137.

[30] F. E. Harrell, K.L. Lee, D.B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistic in Medecine*, vol. 15, no. 4, pp. 361–387, 1996, doi:10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4.

[31] P. Peduzzi, J. Concato, E. Kemper, T.R. Holford, A.R. Feinstein, "A simulation study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, vol. 49, no. 12, pp. 1373–1379, 1996, doi:10.1016/s0895-4356(96)00236-3.

[32] L.E. Eberly, "Multiple linear regression," in *Topics in Biostatistics*, W. T. Ambrosius, Totowa: Humana Press, 2007, pp. 165–187, doi: 10.1007/978-1-59745-530-5_9.

[33] N. M. Jakopin, A. Klein, "Determinants of broadband internet access take-up: Country level drivers," *Journal of Policy, Regulation and Strategy for Telecommunications, Information and Media*, vol. 13, no. 5, pp. 29–47, 2011, doi: 10.1108/14636691111160626.

[34] N. Kock, L. Gaskins, "The mediating role of voice and accountability in the relationship between Internet diffusion and government corruption in Latin America and Sub-Saharan Africa," *Information Technology for Development*, vol. 20, no. 1, pp. 23–43, 2014, doi: 10.1080/02681102.2013.832129.

[35] P.F. Musa, P. Meso, V.W. Mbarika, "Toward sustainable adoption of technologies for human development in sub-Saharan Africa: Precursors, diagnostics, and prescriptions," *Communications of the Association for Information Systems,* vol. 15, no. 33, pp. 592–608, 2005, doi:10.17705/1CAIS.01533.

[36] *S.* Creese, W.H. Dutton, P. Esteve-González, R. Shillair, "Cybersecurity capacity-building: cross-national benefits and international divides," *Journal of Cyber Policy*, vol. 6, no. 2, pp. 214–235, 2021, doi: 10.1080/23738871.2021.1979617.

[37] L. Kano, E. W. Tsang, H. W. C. Yeung, "Global value chains: A review of the multi-disciplinary literature," *Journal of international business studies*, vol. 51, no.4, pp. 577–622, 2020, doi: 10.1057/s41267-020-00304-2.

[38] K. Vu, K. Hartley, A. Kankanhalli, "Predictors of cloud computing adoption: A cross-country study," *Telematics and Informatics*, vol. 52, no. 101426, 2020, doi: 10.1016/j.tele.2020.101426.

[39] M. M. Alam, M. W. Murad, "The impacts of economic growth, trade openness and technological progress on renewable energy use in organization for economic co-operation and development countries," *Renewable Energy*, vol. 145, pp. 382–390, 2020, doi: 10.1016/j.renene.2019.06.054.

[40] N. Ameen, R. Willis, M.H. Shah, "An examination of the gender gap in smartphone adoption and use in Arab countries: A cross-national study," *Computers in Human Behavior*, vol. 89, pp. 148–162, 2018, doi: 10.1016/j.chb.2018.07.045.

[41] V. Dutot, V. Bhatiasevi, N. Bellallahom, "Applying the technology acceptance model in a three-countries study of smartwatch adoption," *The Journal of High Technology Management Research*, vol. 30, no.1, pp. 1–14, 2019, doi: 10.1016/j.hitech.2019.02.001.

[42] H. Edquist, P. Goodridge, J. Haskel, "The Internet of Things and economic growth in a panel of countries," *Economics of Innovation and New Technology*, vol. 30, no. 3, pp. 262–283, 2021, doi: 10.1080/10438599.2019.1695941.

[43] N. Terzi, "The impact of e-commerce on international trade and employment," *Encyclopedia of e-commerce development, implementation, and management,* (IGI Global), pp. 2271–2287, 2016.

[44] D. J. Leu, "Our children's future: Changing the focus of literacy and literacy instruction," *The Reading Teacher*, vol. 53, no. 5, pp. 424, 2000.

[45] N. A. G. Arachchilage, S. Love, "Security awareness of computer users: A phishing threat avoidance perspective," *Computers in Human Behavior*, vol. 38, pp.304–312, 2014, doi: 10.1016/j.chb.2014.05.046.

[46] S. Karunakaran, K. Thomas, E. Bursztein, O. Comanescu, "Data breaches: User comprehension, expectations, and concerns with handling exposed data," Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018), Baltimore, 2018, pp. 217–234, 2018.

[47] Government of Canada. "Breach of Security Safeguards Regulations," 2018. [Online]. Available: https://gazette.gc.ca/rp-pr/p2/2018/2018-04-18/html/sor-dors64-eng.html. [Accessed: July 28, 2023].

Appendix A

Description of password performance by country (N=200).

Country	Mean time to crack in seconds	Minimum	Maximum	Number of users in the list	% of passwords cracked in less than a minute
Australia	59767.98	0	10713600	3083341	79
Austria	1026148.21	0	96422400	695307	68.5
Belgium	22398.15	0	1036800	729661	59.5
Brazil	16137346.42	0	3214080000	4943358	55.5
Canada	4998296.04	0	996364800	5277926	81
Chile	7088718.69	0	1221350400	846354	49.5
China	105039.75	0	2332800	14739683	62.5
Colombia	571740.54	0	96422400	1379631	68.5
Czech R.	81197.92	0	10713600	2288530	65.5
Denmark	489720.11	0	96422400	862571	63.5
Estonia	72966.45	0	10713600	169656	40.5
Finland	24447.04	0	1036800	268236	44
France	67423.91	0	10713600	16160255	54.5
Germany	542675.26	0	96422400	28364318	75.5
Greece	811402.15	0	160704000	861187	80
Hungary	3819.54	0	259200	1159682	48.5
India	1105793.28	0	96422400	8186249	41
Indonesia	43517736.92	0	3214080000	3223828	49
Ireland	2662.44	0	86400	590381	69.5
Israel	10124.02	0	1036800	793908	92
Italy	94203.67	0	10713600	14030845	46.5
Japan	8406.08	0	1036800	1906700	67.5
Korea	949.88	0	86400	910432	84.5
Latvia	23643.52	0	1036800	181072	55

Country	Mean time to crack in seconds	Minimum	Maximum	Number of users in the list	% of passwords cracked in less than a minute
Lithuania	125648.91	0	10713600	406310	40.5
Malaysia	8881.79	0	1036800	1359725	69
Mexico	497790.64	0	96422400	2162221	65.5
Netherlands	767603.60	0	128563200	1636625	56
New Zealand	191147.84	0	32140800	1367054	64
Nigeria	52514.64	0	5356800	757126	40.5
Norway	7633.67	0	1036800	528173	64.5
Philippines	27538.36	0	1036800	2750631	62.5
Poland	10237.31	0	1036800	4412538	46
Portugal	6186350.94	0	996364800	2282038	41.5
Romania	34072.51	0	1036800	1509270	46
Russia	140715.11	0	26784000	146837497	84.5
Saudi Arabia	561965.62	0	96422400	547759	58.5
Slovak Republic	7275.04	0	1036800	702289	51
South Africa	2749.12	0	86400	609061	61.5
Spain	5676283.41	0	996364800	5493452	58.5
Sweden	5091.35	0	172800	1194218	62
Switzerland	82651.48	0	10713600	657863	77
Thailand	560172.86	0	96422400	2055344	65
Total	2082684.37	0	3214080000	3944162	40.5
Turkey	3809402.54	0	514252800	1829898	77
Ukraine	75958.79	0	10713600	529433	78
United Arab Emirates	1515.82	0	86400	7440559	80.5
United Kingdom	6120.43	0	1036800	31229262	84.5
United States	759.68	0	86400	6026634	12
Vietnam	6344824.74	0	996364800	3083341	79

Appendix B

List of the variables that have been tested before determining the final model.

Female participation in workforce (2019)
Freedom of press (2019)
Legal framework's adaptability to digital business models (2019)
Digital skills (2019)
Digital Adoption Index (2016)
DAI Business Sub-index (2016)
DAI People Sub-index (2016)
DAI Government Sub-index (2016)
Number of secured servers (2020)
Mobile cellular subscription (2019)
Voice and accountability (2020)
Political stability (2020)
Government effectiveness (2020)
Regulatory quality (2020)
Rule of law (2020)
Control of corruption (2020)
National Cybersecurity Index (2020)
Global Cybersecurity Index (2020)
Basel AML Index (2020)
Cybersecurity Exposure Index (2020)
Cyber Legislation Rating (2020)
Cyber-Safety Score (2020)
GDP per Capita (2020)
Data breaches (2021)
Internet Users (2020)
IQ (2022)
Literacy (2022)
Education (2022)

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Trust Framework on Exploitation of Humans as the Weakest Link in Cybersecurity

Daudi Morice | Department of Computing Science Studies, Faculty of Science and Technology, Mzumbe University, United Republic of Tanzania,
ORCID: 0000-0001-7907-427X

Abstract

The significance of cybersecurity is increasing in our daily digital lives. The reason for this rise is that human interactions take place in computer-mediated environments, or cyberspace, where physical cues from face-to-face interactions are either absent or very minimal. Computer users are becoming increasingly susceptible to cyberattacks as a result of human interactions in cyberspace. Understanding how cybercriminals exploit the human trust, the weakest link in cybersecurity is relevant because cybercriminals focus on attacking the human psychology of trust rather than technical-based controls. To this end, the present paper develops a trust framework on exploitation of humans as the weakest link in cybersecurity. The framework is established by linking the human psychology of trust and techniques used by cybercriminals in deceiving and manipulating users of computer systems. The framework is validated by demonstrating its application using a case study employing real data. Findings show that cybercriminals exploit human trust based on trust development processes and bases of trust, either creating (falsified) expectations or a relationship history to lure the victim in. Furthermore, it is revealed that technical-based controls cannot provide effective safeguards to prevent manipulation of the human psychology of trust.

Keywords
cybersecurity, human layer, weakest link, trust, trust framework, human trust exploitation

Received: 29.06.2023

Accepted: 16.10.2023

Published: 31.12.2023

Cite this article as:
D. Morice "Trust Framework on Exploitation of Humans as the Weakest Link in Cybersecurity," ACIG, vol. 2, no. 1, 2023, DOI: 10.60097/ACIG/162867.

Corresponding author:
Daudi Morice, Department of Computing Science Studies, Faculty of Science and Technology, Mzumbe University, United Republic of Tanzania, ORCID: 0000-0001-7907-427X
E-MAIL: morice.daudi@mu.ac.tz; dmorice@mzumbe.ac.tz

1. Introduction

Cyberattacks have evolved in many forms and stages. From attacking computers and computer networks, today's cyberattacks also target human beings. This progression, which Schneier [1] refers to as waves of attacks, consists of physical attacks, attacks that target vulnerabilities, and semantic attacks. According to the author, the first wave comprises attacks against computers, wires, and electronics, the second targets vulnerabilities in software products, cryptographic algorithms, protocols, and denial-of-service, and the third targets how humans assign meaning to content. State officials in charge of upholding the law, such as the police, have documented numerous incidences, particularly for the third wave. For example, between 2017 and 2020, 19,530 cybercrime incidents were reported to police in Tanzanian [2–5]. Many of these incidents were committed through social engineering techniques relating to how humans assign meaning to content. Such ever-increasing cybercrimes are emphasized by Schneier [1], who posits that semantic attacks will become more serious than physical or even syntactic attacks in future and that dismissing them using cryptographic measures will be difficult. Given that humans are the weakest link in computer information system components, semantic attacks target people more than other components. That shift in target is partly attributed to the relative strength of technical-based controls in cybersecurity. Technical-based controls are difficult to crack compared to human being psychology, which is easy to manipulate.

Technical controls are built on the triad of Confidentiality, Integrity, and Availability (CIA). These principles are widely used to ensure the security of computer resources. Despite their advantages, CIA concepts have several drawbacks. First, the CIA framework focuses on isolating legitimate from illegitimate users, granting legitimate users full access to computer resources, which that user is privileged to access. Once users are considered fair and granted access, CIA primitives provide the least control over actions users can perform. Second, CIA principles rely on algorithms developed based on historically conceptualised cybersecurity incidents. That history dependence implies that new incidents that have not been conceived are difficult to control and manage. Given these restrictions and a rise in attacks on humans compared to cryptograph-based methods, there is a need for human-centric complementary defence. That need is imperative because technology is not the only way to address information security risks [6]. Furthermore, customers and organisational insiders make information security challenging [7], as their misbehaviour can directly or indirectly lead to cybercrime. Since

most amateurs attack machines while professionals target people, cybersecurity solutions must now target humans more [1] than ever before. Creating human-centric cybersecurity solutions necessitates collaboration between industry and academia to brainstorm from an alternative perspective. One of those perspectives is trust, a human component many cybercriminals exploit. The critical question may be, "How do humans come to trust cybercriminals?"

Humans play trusting roles in cybersecurity at a moment when technical-based controls fail to detect and prevent cyberattacks. One area contributing to cybercrime attacks involves trust between computer users and cybercriminals. Cybercriminals prefer to exploit people's trust rather than technology since it is easier to exploit their natural inclination to trust [8]. It is also simple to deceive people if you can gain their trust. For this reason, cybersecurity requires managerial efforts on top of technical-based controls to combat cybercrime.

Cybercrime occurs at many layers, much like those of Open Systems Interconnection (OSI) and TCP/IP models. These cybersecurity layers include mission critical assets, data security, application security, endpoint security, network security, perimeter security, and the human layer [9, 10]. Among them, human is the weakest and most vulnerable layer. The co-existence of these layers implies that technological and management or policy-based control must be implemented nearly concurrently. Though they coexist, the human layer depends more on policy-based controls than technical-based solutions. This is due to the fact that human trust behaviours manifest in reasoned decisions and actions that are not part of coded algorithms. Rather, trust behaviour results from an individual's mental ability to either accept or reject cooperation with a counterpart based on the degree of trust the user builds. Trust in humans is attributed to inherent characteristics, which are part of the individual or "given" by the trust-giver, and situational characteristics external to the individual [11].

Many technical solutions have been developed to counteract cyberattacks. However, the number of cyberattacks continues to increase due to inherent constraints in CIA doctrines and a shift in emphasis on exploiting humans as the weakest link. In [12] the authors describe technical and non-technical state-of-the-art protection tools related to everyday online activities. In [13], the authors analyse models of human behaviour that impact data system protection and how systems can be improved and highly secured against any vulnerabilities. These options, which are typical of numerous existing

alternatives, are insufficient on their own. Human beings continue to be the weakest link in cybersecurity, a fact that cybercriminals know and take advantage of by exploiting the human psychology of trust. Meanwhile, algorithms for detecting and preventing human trust exploitation are scarce in the literature. This is confirmed by Shabut et al., who contend that an intelligent tool capable of comprehending cyberattack mechanisms and user behaviours involving assumptions, decision-making, and responses to cyber threats/risks is currently lacking [12]. Alternatively, users must be aware of how cybercriminals exploit human trust instead of depending only on coded algorithms. To this end, the overall objective of the present paper is to examine in detail how cybercriminals exploit human trust. The paper contributes by formulating a trust framework on the exploitation of humans as the weakest link in cybersecurity. It answers the two following research questions:

- How do cybercriminals exploit human trust, the most vulnerable link in cybersecurity?
- How can the development of a trust framework on the exploitation of humans as the weakest link in cybersecurity be beneficial?

The paper contributes to helping individuals and organisations know and gain awareness of trust development processes and bases of trust that cybercriminals employ to manipulate and deceive users of digital systems and gadgets. That awareness enables individuals and organisations to detect, react and prevent attacks on human trust, leaving them better equipped to recognise, respond to, and stop such attacks.

2. Trust and Cybersecurity in the Human Layer

The present section covers discussions on trust and cybersecurity in the human layer. The human layer of cybersecurity is covered in subsection 2.1. Common cyberattacks affecting individuals and organisations are covered in more detail in subsection 2.2. Subsection 2.3 concludes the discussion by presenting a thorough analysis of trust in computer-mediated environments.

2.1. Cybersecurity in the Human Layer

The human layer of cybersecurity is part of cyberspace, a time-dependent set of interconnected information systems and human users that interact with these systems [14]. It is in this space,

cyberspace, where cybercrime occurs. Cybercrime can essentially be regarded as any crime (traditional or new) that can be conducted or enabled through digital technologies [15]. Such crimes must be controlled and prevented to safeguard data, information systems, and users. Consequently, the act of detecting, reacting and preventing cybercrime is referred to as cybersecurity. The authors in [16] define cybersecurity as the organisation and collection of resources, processes, and structures used to protect cyberspace and cyberspace-enabled systems from occurrences that misalign de jure from de facto property rights. With that brief overview, the next paragraph contextualises cybersecurity in the human layer.

Today's cryptographic magic wands of "digital signatures", "authentication", or "integrity" [1] are not the ultimate protective mechanism to rely on. These cryptographic techniques can barely identify most lies that manipulate the human psychology of trust. Cybercriminals have a long history of taking advantage of the psychological needs and vulnerabilities of people in a variety of ways, including the human need for love and affection, our fundamental desire to be trustworthy and helpful, and the many biases that influence security decision-making [17]. Another form relates to perfect knowledge of what people consider most important [15]. These outlined techniques are sources of human weaknesses that cybercriminals employ as weapons to exploit individuals and organisations.

2.2. Cyberattacks in Cyberspace

Cybersecurity incidents impact individuals and organisations worldwide, causing harm to social and economic values. They involve malware, password theft, traffic interception, phishing, denial-of-service, cross-site (xss), zero-day exploits, social engineering, and crypto-jacking. However, other types of cybercrime, such as terrorism, cyber warfare, cyber espionage, and cyberbullying, are also emerging. All of these threats originate in the digital environment in networked and non-networked computer systems.

Cybercrime threats affect our everyday life, from financial transactions to social interactions. For example, reports from the Inspector General of Police show that over four years, 19,550 incidents of cybercrime were reported in Tanzania [2–5] (Table 1).

Table 1. Selected cybersecurity incidents in four years in Tanzania [2–5].

Type of Cyber Incident	Year				Cumulative Sum
	2017	2018	2019	2020	
Theft	2,568	4,310	2,408	2,963	12,249
Death Threats	447	851	409	490	2,197
Insults	489	757	287	385	1,918
Threats	51	61	43	254	409
Misuse of the Internet	11	374	19	0	404
Trusted Theft	81	203	31	65	380
Fraud	48	282	1	21	352
ATM Theft	171	20	97	53	341
Financial Fraud	52	40	83	89	264
Forgery	78	112	21	32	243
Attempted Financial Fraud	17	2	10	111	140
Network and System Intrusion	26	15	0	54	95

According to that report, theft, death threats, and insults are the major cybercrime incidents being reported to the police. Such statistics correspond to a remark emphasised in The Citizen that theft via mobile money transactions, abusive language, and theft of information shared on various cyber platforms are frequently committed crimes [18]. It is estimated that 91% of cybercrime cases go unreported to the police [19], suggesting that 19,530 cybercrime recorded incidents may reflect underreporting of cases by organisations and individuals.

Some national and international organisations are already implementing strategies to fight cybercrime. For instance, AFRIPOL [20]has been fighting cybercrime by raising awareness, reinforcing policy and legislation to fight cyber criminals, and implementing technologies to support cyber-defence. Similar measures to combat cybercrime are also recommended in other literature sources. The Tanzania Cybersecurity Report of 2016 recommends improving internet user education [19] in fighting cybercrime. Educating users also means raising

their cybersecurity awareness, which is critical, especially for organisations that have many employees. Since research shows that over 80% cases of system-related fraud and theft in 2016 were perpetrated by employees and other insiders [19], training employees on proper internet use and how to fight cyberattacks is essential. Moreover, it is indispensable to extend training on cybersecurity awareness to individuals in the local community.

Conversely, cybercriminals play on human psychology to manipulate users, and gain or guess their access credentials. Evidence of this claim is featured in weekly reports[1] released by [21] TZ-CERT in Tanzania. TZ-CERT studies cyberattack patterns by setting up a honeypot. The honeypot is a network-attached system set up as a decoy to lure cyber attackers and detect, deflect or study hacking attempts meant to gain unauthorised access to information systems. The resulting information helps to guide users of computer systems in many ways, including how to prevent cyberattacks. According to TZ-CERT reports, which are analysed in Figure 1, cybercriminals use human psychological heuristics – based on the human inclination to use default, simple, or common access credentials – to guess usernames and passwords.

1 —— Seven reports released from 5 to 17 July 2023.

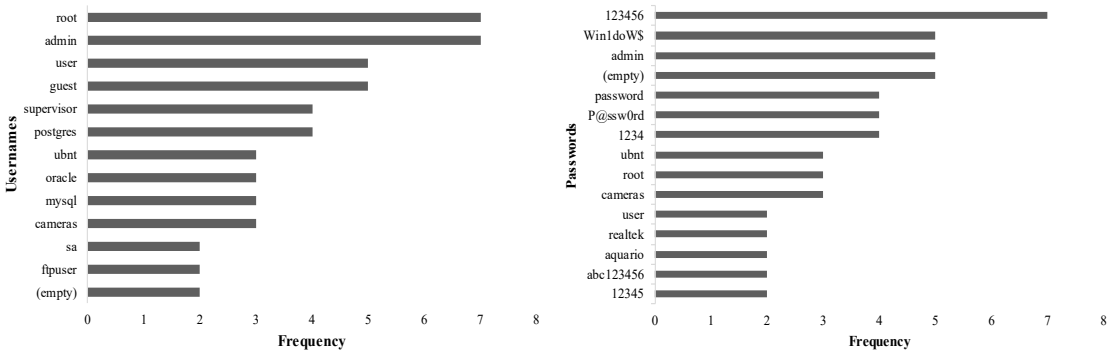

Figure 1. Common usernames and passwords used by cybercriminals.

Usernames such as "*root*", "*admin*", "*user*", "*guest*", "*supervisor*", and "*postgres*", and passwords such as "*123456*", "*win1doW$*", "*admin*", "*(empty)*" and "*password*" were common. These usernames and passwords are easy to remember, hence their prevalence. The access credentials in Figure 1 presents the psychological behaviour of many computer system users when choosing usernames and passwords.

2.3. Trust in Computer-Mediated Environment

The current section discusses trust in a computer-mediated setting, investigating the risk viewpoint of trust (subsection 2.3.1), as well as behavioural control in interactive systems (subsection 2.3.2). It concludes with a discussion of bases of trust (subsection 2.3.3).

2.3.1. Risk Perspective on Trust in Relationships

For an exchange to be completed, two parties, a trustor and a trustee, must be engaged. A trustor is an entity that develops a degree of reliance on another object and accepts being vulnerable to the possible actions of that other object [22]. Similarly, the trustee is the party in whom the trust resides, who can exploit the trustor's vulnerabilities [23]. The trustor is the party that puts its expectations in the other party, while the trustee is the party in which that expectation resides. While many definitions of trust exist, the present paper adopts the following definition: Trust is a level of confidence a trustor develops in a trustee based on the expectation that the trustee will perform a particular action necessary to the trustor [22].

Trustors (humans and other objects) live in a partially unpredictable world because of their limited ability to know trustees (surroundings). A trustor is neither in total ignorance nor fully informed concerning a trustee. Under complete ignorance of information, decisions to trust are risky; thus, a transaction of trust should be avoided. Trust becomes meaningless if the trustor has complete information (full rationality) because one can rationally predict before acting. However, in practice, total ignorance and rationality are unrealistic. Humans interact and collaborate in a bounded world where risks are neither fully predictable nor ignored. This situation of partial predictability exposes human beings to a risky world, thus creating a need for trust.

2.3.2. Behavioural Control in Interactive Systems

In recent years, humans, physical robots, bots, and organisations have started to coevolve and interact. These convergent interactions are managed under the security, institutional, and social control approaches. In subsequent paragraphs, attention is drawn to the fact that the term "agent" refers to people, physical robots, bots, and organisations.

a. **Security Control.** Security is a binary control mechanism that attempts to distinguish between agents' contextual behaviour. The security principles offer a sphere of compliant agents while

creating a wall that prevents non-compliant ones. Norms restricting interactions under security control are pre-defined as rules and regulations before being reinforced. Once defined, those rules and regulations are reinforced by scrutinising each agent to verify if it complies. Therefore, security control deals with a binary choice between yes and no, legitimate or illegitimate, acceptance or sanctions [24], authentic or unauthentic, and approval or disapproval. One disadvantage of security controls, particularly in computer systems, is that once an attacker is accepted as genuine and authorised, there is no cap on the privileges it can exercise, rendering it free to carry out (malicious) actions without any extra constraints.

b. **Institutional Control**. Institutional control entails a central authority to monitor, regulate, or enforce the acts taken by agents, and punish those agents who engage in undesired behaviours [25]. For example, police, judicial systems, regulatory bodies, and companies use institutional control to influence the behaviour of individuals and organisations [26] designing for trust in mediated interactions has become a key concern for researchers in human computer interaction (HCI). This form of formal control goes through articulated procedures specifying rewards and punishment. For instance, communication regulatory bodies in various nations and areas monitor online transactions and can testify in court and to the police about cybercrime charges reported.

c. **Social Control**. By enforcing social norms, social control regulates agent interaction in systems. A social norm sanction refers to societal approval or disapproval, which is difficult to determine in advance [27]. Social norms are enforced through social sanctions, which create a range of unpleasant emotional states in those who have violated them [28]. Social control mechanisms don't deny the existence of malicious entities but attempt to avoid interaction with them [29]. In this approach, agents can punish non-desirable behaviours, for instance, by not selecting certain partners [25].

2.3.3. Bases of Trust in Inter-Personal and Business Relationships

Trust is derived from various sources or bases in both personal and business relationships. According to [30], trust can be based on mechanisms of deterrence, cognition, affection, and calculus, as well as formal and informal institutions. Subsequent

paragraphs discuss such bases of trust and how they can apply specifically in cybersecurity incidents (Tab. 2). These bases are adapted from [31].

a. **Calculus-based trust.** Calculus-based trust plays a major role, especially at the beginning of a relationship. As a form of trust-building process, calculus-based trust is founded on: calculating the rewards and costs of committing a transaction, thereby developing confidence that the trustee's behaviour can be predicted, and assessing the trustee's ability to fulfil its promises [32]. Calculus-based trust may be assessed rationally based on credible information sources (reputation, certification) about the trustee. It depends on a rational choice that involves characteristics of interactions founded on economic exchange [33], and deals with factors such as relationship economics and the dynamic capabilities of partners [34]. As the weakest link, humans have to calculate the cost and reward of cooperating based on the level of trust they place in the cybercriminal. Under calculus-based trust, some cybercriminals opt to offer falsified economic benefits, which later turn out to be deception of a victim (computer user).

b. **Deterrence-based trust.** Deterrence occurs when the potential costs of breaking up a relationship outweigh the immediate advantage of acting distrustfully [35]. Deterrence-based trust mechanisms consist of evaluating the advantages and costs of continuing in the relationship, the rewards and costs of cheating on the relationship, and the benefits and costs of quitting the relationship [36].

Table 2. Bases of trust in the human layer of cybersecurity (adapted from [31]).

Basis of trust	Foundation	Description
Process-based trust	Tied to past or expected exchange	Developed based on past or repeated exchanges between cybercriminals and target.
Institution-based trust	Tied to formal social structure, broader societal institutions	Attributes of a person or firm, or an intermediary mechanism shape the possibility for trust to arise.
Deterrence-based trust	Fear of consequences	Behavioural consistency is constrained by the potential costs of discontinuing the relationship.

Basis of trust	Foundation	Description
Competence trust	Based on the partner's competency	An actor predicts others' abilities and expectations of whether they will perform roles competently.
Calculus-based trust	Based on rational choice	Related to the perception of benefit from the relationship.
Relational trust	Tied to repeated interaction	From repeated interaction, the parties obtain information and experience that engenders trust.
Knowledge--based trust	Based on a sufficient understanding of the other party	Prediction of the other party's behaviour based on the history of the relationship.
Identification--based trust	One party has fully internalised the other's preferences	Understanding others' wants. This is the highest level of trust.

Cybercriminals sow fear by threatening users of computer systems to meet falsified demands, which appear to be genuine.

c. **Institutional-based trust.** Institutional trust is tied to formal social structure and broader social institutions. According to [31], the conditions for institutional-based trust are shaped by personal or firm-specific attributes or intermediary mechanisms. Taking advantage of institutional-based trust, cybercriminals impersonate the employees of a particular company, earning the trust of a computer user, who can then be exploited.

d. **Relational trust.** Relational trust refers to the extent to which one feels a personal attachment to the other party and wants to do good for the other party, regardless of egocentric profit motives [37]. The key to relational trust is that one party empathises with the other party and wants to help them for altruistic reasons [37]. Variations of relational trust include normative trust, good will trust, affect-based trust, companion trust, fairness trust, and identification trust [37]. The human psychology of trust is exploited by cyber attackers who understand human perceptions of kindness and unselfishness well, which exposes cybercrime victims to subsequent consequences.

e. **Identification-based trust.** Identification-based trust involves identification with the other's desires and intentions, i.e. trust

exists because one party effectively understands and appreciates the other's wants [38]. This mutual understanding is developed so that each party can effectively act for the other. Identification-based trust is grounded in deep knowledge of the partner's desires and intentions [39]. Identification-based trust can be used to exploit human trust when the trustor and trustee understand each other, as well as when the trustor and trustee have common intentions and desires, e.g. trusting someone to use your electronic gadget. It can also include allowing someone to use your account to access electronic systems, as well as intentionally sharing your credentials with a third party.

f. Knowledge-based trust. Knowledge-based trust is grounded in the other's predictability, or sufficient knowledge that allows the other's behaviour to be anticipated, and relies on information rather than deterrence [38]. Knowledge-based trust develops over time through a track record of interactions that enable both parties to build generalised expectations about each other's behaviour [39]. By being predictable, cybercrime victims are exposed to the actions of the cyber attacker because the cyber attacker knows all the possible means to deceive and manipulate the target, as well as how the target usually responds. Generally, if a person is rationally predictable, that person can be taken advantage of.

2.4. Trust Development Processes

Trust develops from relationship history and subsequent expectations processes. Trust developed from relationship history usually results from the past or previous relationships with people, or other entities or objects [31]. Through relationship history, trust develops based on how parties have previously interacted and the experiences they have gained from one another. When parties have had no previous direct interactions, reference from a third party is usually used to infer the development of trust. Inference is used because, under relationship history, trust develops through interactions with partners that we meet directly or indirectly. Examples of bases of trust that develop from relationship history include knowledge-based, relational, and process-based. Process-based trust production emphasises that past exchanges, whether through reputation or direct experience, lead to a perception of trust in the counterpart [40].

The second process of trust development involves future expectations. Humans may trust the other party by relying on what

they expect to gain after a trust transaction has been performed. Thus, trust formed in this way is usually based on a consideration of the benefits and costs related to a particular relationship [31]. Deterrence-based and calculus-based trust, for example, both rely on future expectations. One party may choose to trust another party after calculating the cost and benefits of an existing relationship. It may also opt to trust because of fear generated by another party.

3. Methodology

The present paper adopts the methodology in [41], developing a theoretical framework that predicts correlations between trust and human behaviours in the cybersecurity layer. In accordance with this methodology, the scope of this paper comprises an analysis of common cyberattacks encountered by users of computer systems, as well as theoretical foundations in trust and cybersecurity. In the former, cybercrime cases reported to police in Tanzania are analysed to indicate how widespread the problem is. Next, a discussion on systems used to control behaviours in interactive systems that fall under face-to-face and computer-mediated environments is presented. Bases of trust that can be used to manipulate human trust are also analysed in detail, and the ways in which trust is employed by cybercriminals to exploit computer users are presented. Generally, most of the discussion is centred on humans as the weakest link in the cybersecurity layer, where human trust is primarily exploited.

Subsequently, the study develops a trust framework to describe how easily human trust can be exploited compared to technical-based controls. This development reveals how deceptive and manipulative attacks on the human psychology of trust go undetected by considering technical and non-technical controls. The study uses data from Tanzania to validate and show the practicality of the framework, which comprise real cases of cybersecurity incidents that were directly observed by the researcher. Secondary data, or cybersecurity incidents reported in the literature originating in Tanzania, are also used.

Additionally, the following are considered during validation and demonstrative application of the trust framework. First, each reported cybercrime incident is explicitly linked to a specific trust formation process. Second, such cybercrime incidents are further linked to bases of trust. This linkage serves to demonstrate how human trust is exploited differently in various circumstances.

Moreover, two issues are taken into consideration throughout the validation and demonstration of application of the trust framework. First, a specific trust formation process is explicitly linked to every cybercrime incident that has been reported. Second, a specific basis of trust is further connected to each cybercrime incident. This connection helps to show how different circumstances lead to different forms of exploitation of human trust.

4. Cybersecurity Trust Framework in the Human Layer

This section details the fundamental structure of trust in the human layer of cybersecurity, which comprises a trustee (cyber attacker) who is regarded as a cybercriminal, and a trustor, usually the end user who is commonly referred to as a cybercrime victim. The cybercriminal and cybercrime victim are the main actors who usually engage in communication.

For a cybercrime incident to occur there must be virtual and occasionally physical interactions between the cybercriminal and victim. Presumably, the victim is protected by technical-based controls, but also policy-based controls, which the victim has to exercise. With those two defences in place, cybercriminals may choose to attack technical protective mechanisms or the human psychology of trust. The latter is the weakest link in the cybersecurity layer. Attacking human psychology of trust is easy to achieve and requires less effort. Most cyberattacks on technological controls can be mitigated by technical countermeasures such as solutions based on encryption, firewall, antivirus, and access control techniques. Insofar as the human layer is excessively exploited, the current framework focuses on exploitation of human trust.

The exchange of exploitative cues between trustor and trustee are hard to detect and prevent by using algorithms inherent in technical-based control systems that protect computer resources (Fig. 2). This is because a criminal communicates as if s(he) is a legitimate user. This may involve forged identities (such as gadgets, or authorised or unauthorised communication channels) so effectively that computer equipment scrutinising the signals travelling through it uncovers no evidence of susceptibility. These devices are rendered incapable of detecting vulnerability even though they typically perform their protective tasks well based on the functions for which they were built and developed. For example, if a cybercriminal communicates lies via voice or text, the algorithms in those computer

devices are unlikely to detect it. Computers rarely detect malevolent intent when a user obtains authorisation and is provided access to systems.

Assume that the cybercriminal wishes to exploit the victim's human psychology of trust (Fig. 2). The cybercriminal must choose the basis of trust to use, depending on whether the cybercriminal and victim have previously interacted. If there has been prior interaction, the cybercriminal will employ bases of trust whose development relies on a relationship history. Otherwise, they will resort to bases of trust that can be developed in anticipation of future expectations.

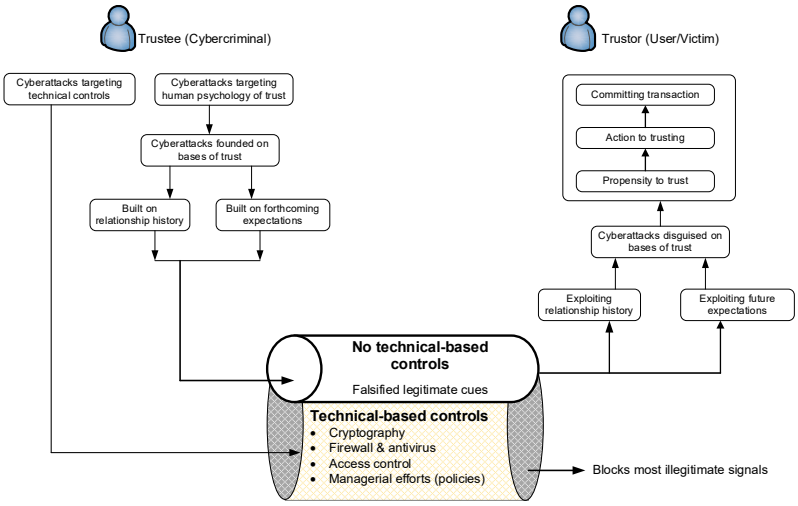

Figure 2. Cybersecurity trust framework in the context of the human layer.

Depending on prevailing circumstances and choice of method or technique, the cybercriminal can use one or multiple bases of trust to deceive and manipulate the victim. Those bases of trust are described fully in subsection 2.3.3. For example, the cyber attacker may use deterrence-based trust to threaten the victim, offering a reward if its demands are met, and punishment if they are not (Table 3). The choice of basis of trust is associated with the trust development process with some using both trust development processes, such as characteristic-based trust and competence-based trust. Other bases of trust employ one trust development process, for example, calculus-based trust or knowledge-based trust.

Table 3. Bases of trust and their development processes.

Bases of Trust	Trust Development Process	Brief Descriptions
Characteristic-based trust	• Relationship history • Future expectations	Provides background to develop a mutual understanding [31]
Institution-based trust	Relationship history	Attributes of a person or firm, or an intermediary mechanism shape the possibility for trust to arise [31].
Deterrence-based trust	Future expectations	Out of fear, the cybercrime victim meets the cyber attacker's demands, creating future expectations of avoiding harm
Competence trust	• Relationship history • Future expectations	A cybercrime victim develops expectations based on the capability (competence) of the cyber attacker
Calculus-based trust	Future expectations	Weighing the benefit and losses of a relationship
Knowledge-based trust	Relationship history	Behavioural predictability based on available information
Identification-based trust/ Relational trust	Relationship history	A mutual understanding of desires and intentions, as well as the need to feel unselfish and desire to help

The signal/cues communicated by the cybercriminal pass via technical-based solutions undetected. They remain undiscovered because they are (falsified) valid cues in disguise. As a result, the cybercrime victim responds to the cybercriminal, assuming that the signals sent are legitimate.

After the cybercrime attacker gains the victim's trust, it engages the victim in acts that appear innocuous at first. At that stage, the trustor (victim) develops expectations in the fictitious transaction in the hope of obtaining prospects of the assumed agreement. That stage is known as action to trusting. With that expectation in mind, the victim enters into a transaction after trusting the cybercriminal, performing action(s) that fulfil a promise made between

the cybercriminal and victim. Following that transaction, the victim, in accordance with trust standards, compares expectations to the resulting outcome, which may become apparent immediately or later. Furthermore, victims may not realise they have been attacked, depending on the severity of the cyberattack. Some examples include the theft of email passwords where the attacker has no intention of blocking the user's account. In general, the outcome will differ from what the user expected.

5. Case Illustration

The present section provides an illustrative application of the trust framework in the context of the weakest link in cybersecurity. Cybersecurity incidents extracted from real scenarios are presented in Table 4. Some were collected from literature, while others were encountered by the author on various occasions.

In case 1, the cybercriminal uses a mobile phone to deceive users by pretending to be a landlord. The cybercriminal broadcasts information via a Short-Message-Service (SMS) to multiple mobile phone users simultaneously. In that attempt, at least one user may have rented a house from another real landlord, and receives a message concerning rent to be paid. When the SMS is received, the tenant may get confused about whether the sender is the actual landlord or not. The tenant is manipulated further by being directed to pay the rent to a mobile number provided in the SMS.[2] In this situation, the requested amount is expected to be transferred using a mobile money service.

Essentially, the cybercriminal acts as if there is an existing relationship with a victim (landlord-tenant relationship), thereby building trust through a falsified relationship history as a trust development process. The cybercriminal exploits the tenant's trust by employing identification-based trust, the highest psychological tool, to manipulate the tenant into understanding the other side's desire. The cybercriminal has also weaponised relational trust, in which kindness and unselfishness are core components. Overall, the cybercriminal operates on the assumption that there is an agreement on rent payment, taking advantage of the landlord's desire to obtain and the tenant's intention to pay rent.

2 — It should be noted that in some African countries, including Tanzania, mobile phones are used to send and receive money in addition to paying various bills via a service known as mobile money. Mobile phones are used to carry out financial transactions at a country and even regional level.

Table 4. Cybersecurity incidents committed through the exploitation of human trust.

Case No	Incident	Trust Development Process	Bases of Trust Used
1	I'm your landlord. My current number is unreachable. Send the rent through this number +255 (*number withheld*).	(Falsified) relationship history	• Identification-based trust • Relational trust
2	Please get in touch with us as soon as you can; your child is extremely ill. Teacher.	• Future expectations • (Falsified) Relationship history	Deterrence-based trust
3	After unexpectedly collapsing at school, your son was brought to the hospital. Send money right away for medical care.	• Future expectation • (Falsified) Relationship history	Deterrence-based trust
4	Don't call; the phone's speaker is broken; instead, send the money to this number +255 (*number withheld*).	(Falsified) Relationship history	• Identification-based trust • Relational trust
5	This is the Revenue Authority office. Why don't you use an electronic fiscal device (EFD) when conducting business? A Tsh 3 million fine is being sent to you immediately.	Future expectations	• Deterrence-based trust • Calculus-based trust
6	You are speaking with someone from the telecom company (*name withheld*); your monthly bonus is TSh 400,000 now. Use a different mobile phone so that we can help you obtain the money.	• Future expectations • Relationship history	• Calculus-based trust • Institutional-based trust
7	This is agent (*name withheld*) from telecom company (*name withheld*). Your mobile money account has insufficient funds. Deposit TSh 500,000 today, then call us back. Otherwise, we are going to close your account.	• Future expectations • (Falsified) Relationship history	• Deterrence-based trust • Calculus-based trust • Institutional-based trust

Case No	Incident	Trust Development Process	Bases of Trust Used
8	I received notification that I had won a customer drawing and was asked to contact a number to learn how to collect my prize. When I called the number, the man instructed me to use 46 as the identification number for prize collection. Then he wanted me to send TSh 60,000 to activate the prize. I sent the money, but when I called the number another time, it was out of service [42].	• Forthcoming expectation • (Falsified) Relationship history	• Institutional-based trust • Calculus-based trust • Characteristics-based trust
9	I received an SMS that appeared to be from M-PESA. The SMS said that I had received TSh 40,000 from a number registered to (*name withheld*). A few minutes after reading the message, someone called and told me he was from Vodacom customer care service. He asked if I had received an SMS that increased my account balance by TSh 40,000. I said I had. Then he asked me to resend the money because it was sent to the wrong account. He told me to send TSh 39,000 to avoid a service charge. When M-PESA replied that the transaction had been successfully completed, I realised my balance had decreased. At that point, I discovered that I had been deceived [42].	Relationship history	• Institutional-based trust • Relational trust

Cybercriminals exploited the tenant's trust psychology because they understand the human perception of kindness and unselfishness. If the tenant cannot sense the deception and use other means to validate whether the received SMS is legitimate, that tenant may end up sending money to a person who is not a real landlord. Such communications pass through digital channels as legitimate cues and are mostly impossible to recognise and filter. Similarly, in case 4, the cybercriminal employs a similar technique to exploit mobile money users.

Cybercriminals also use deterrence-based trust to exploit mobile phone users. In cases 2 and 3, the cybercriminal sends an SMS to targeted parents, informing them that their children are sick. To understand these cases, it should be assumed that some parents send their children to boarding schools. Furthermore, parents are known for their affection and concern for their children; learning that their children are ill can be upsetting and confusing. The cybercriminal (a falsified school teacher or medical doctor) uses deterrence-based trust to introduce a fear that if money is not sent, a child may die from lack of health care. Using deterrence-based trust, the cybercriminal exploits human trust developed through the parent's future expectations of the child's recovery. In addition, the cybercriminal uses institutional-based trust by pretending to be a school teacher, exploiting the parent further to gain trust. This kind of cybercrime employs a common situation in which legitimate teachers and some medical doctors may call parents to obtain additional funds to save a dangerously ill child.

Cases 5, 6, and 7 involve trust building mainly through future expectations and partly through relationship history. In case 5, the cybercriminal communicates via SMS, impersonating an officer of a revenue authority. The falsified officer chooses to create trust with the business owner by setting clear expectations, allowing the owner to believe it is the sole alternative to avoid closure of the business. The business owner is manipulated into believing that if a certain amount is not paid, the revenue authority will close the business. The cybercriminal builds trust through fear (deterrence-based trust) and comparison of the cost and benefit of paying or not paying the falsified fine (calculus-based trust). For cases 6 and 7, the cybercriminal uses mostly future expectations and a (falsified) relationship history to build trust in a mobile money user, relying on calculus-based and institutional-based trust. Through calculus-based trust, a mobile money user compares receiving or losing a bonus (case 6), and making or refusing to make a deposit, and account closure (case 7). Through institutional-based trust, the cybercriminal impersonates an employee of a particular telecom company, gaining more trust from a mobile money user. In case 7, the cybercriminal uses deterrence-based trust to create fear in the mobile money account owner, an agent whose role involves receiving and sending money to mobile money users. The fear is based on the fabricated fact that the account will be cancelled if the owner does not deposit the money. Both cases use relationship history as an additional trust development process. To take advantage of relationship history, cybercriminals impersonate employees of legitimate entities, assuming a legitimate long-term relationship between a mobile money user and a telecom

company. Leveraging that relationship, the mobile money user is further deceived into trusting the cybercriminal.

The last illustration concerns cases 8 and 9, in which both trust development processes are involved. In scenario 8, by setting up future expectations, the cybercriminal communicates that a mobile money user has won a drawing in an attempt to win trust. Falsified winning of the drawing exploits the user's trust as follows: the user compares the benefit and cost (calculus-based trust) of engaging with cybercrime and finally opts to trust because of expectation of winning. The act of trust and committing to sending money is founded on institutional-based trust because the cybercriminal is impersonating an employee of a gambling company. To incorporate relationship history into the trust-building process, the cybercriminal assumes a legitimate long-term relationship that exists between gambling companies and winners. In case 9, the cybercriminal uses relationship history to build trust with a mobile money user. That relationship history is grounded in institutional-based trust because the cybercriminal is impersonating a telecom company employee, exploiting the trust of mobile money users in the company and its employees. To further deepen the trust, the cybercriminal employs relational trust through altruism by asking a mobile money user to return money that was supposedly transferred in error. The mobile money user is exploited by following the cybercriminal's instructions only to find out that their account balance has decreased.

In summary, technical-based controls employed by individuals and organisations rarely detect the above-mentioned techniques of deception and manipulation of human trust since the cues sent to users pass unfiltered via computer-network infrastructure because they are deemed legitimate. Given these limitations, the developed trust framework plays a role in safeguarding users of computer systems.

With respect to the first research question, the present paper argues that cybercriminals exploit human trust based on trust development processes and bases of trust, either creating (falsified) expectations or a relationship history to lure the victim in. Moreover, cybercriminals take advantage of user ignorance of the limitations of technical-based controls. In line with the second research question, the trust framework on exploitation of humans as the weakest link in the cybersecurity layer has many potential benefits and applications. First, the trust framework informs users of computer systems that lies, deception, and manipulation built on human trust can rarely be detected and prevented using technical-based controls. Second, computer system users can identify and stop cybercrime assaults directed at them

by using the trust framework as a guide. Third, people will be less susceptible to cyberattacks if they are aware of the bases of trust that cybercriminals frequently exploit. Finally, computer users will learn and become aware of the way cybercriminals utilise past relationships and future expectations to deceive and carry out cyberattacks.

6. Conclusion

Cybersecurity has become a crucial challenge in this world of digital connectivity because information processing and transfer occurs in cyberspace, which is vulnerable to attacks by many intruders. Recent evidence shows that cyberattacks are increasingly shifting away from technical-based controls to target the human psychology of trust, the weakest link in the cybersecurity layer. Such attacks are linked to how human beings, particularly end users, come to trust cybercriminals. From that viewpoint, the present paper has explored how cybercriminals exploit human trust. In addressing this problem, the paper has established a trust framework to ensure better understanding of how security-based interactions between cybercriminals and victims occur. The framework reveals that trust is a core ingredient in the human – or most vulnerable – layer of cybersecurity. Furthermore, the trust framework indicates that technical-based controls cannot provide effective safeguards to prevent manipulation of the human psychology of trust. Instead, people must protect themselves through greater awareness of cybercrime incidents that are linked to trust. The paper uses real cases to demonstrate the applicability of the trust framework. These scenarios were thoroughly examined, linking them to trust bases and trust development processes. Generally, the sample cases discussed reveal inherent flaws in human trust, which hackers weaponise to deceive computer users.

Despite the extensive discussions presented, this paper has certain limitations and further research may be required to address them. First, this study recognised relationship history and future expectations as trust development processes. In terms of cybersecurity, it is currently unknown which trust development process is more commonly utilised by cybercriminals to exploit human trust. Therefore, future research could investigate common trust development methods employed by cybercriminals. Second, cybercriminals can utilise various bases of trust to deceive and influence computer users. Similarly, greater awareness of which bases of trust cybercriminals employ more often may help organisations and individuals to protect themselves.

References

[1] B. Schneier, "Semantic Attacks: The Third Wave of Network Attacks," *Schneier on Security*, Oct. 15, 2000. [Online]. Available: https://www.schneier.com/crypto-gram/archives/2000/1015.html#1. [Accessed: Feb. 04, 2023].

[2] URT, "Crime and traffic incidents: report January-December 2017," *Dar es Salaam*, 2018. [Online]. Available: https://www.nbs.go.tz/nbs/takwimu/Crime/Crime_Report_January_to_December_2017.pdf. [Accessed: Feb. 04, 2023].

[3] Inspector General of Police, "Takwimu za Hali Ya Uhalifu na Matukio ya Usalama Barabarani Januari – *Desemba* 2018," Dodoma, 2019. [Online]. Available: https://www.nbs.go.tz/index.php/en/census-surveys/crime-statistics. [Accessed: Feb. 04, 2023].

[4] Inspector General of Police. (2020). "Takwimu za Hali Ya Uhalifu na Matukio ya Usalama Barabarani Januari –– Desemba 2019," *Dodoma*. [Online]. Available: https://www.nbs.go.tz/index.php/en/census-surveys/crime-statistics. [Accessed: Feb. 04, 2023].

[5] Inspector General of Police, "Takwimu za Hali Ya Uhalifu na Matukio ya Usalama Barabarani Januari – Desemba 2020," *Dodoma*, 2021. [Online]. Available: https://www.nbs.go.tz/index.php/en/census-surveys/crime-statistics. [Accessed: Feb. 04, 2023].

[6] W. D. Kearney, H. A. Kruger, "Considering the influence of human trust in practical social engineering exercises," in *Proceedings of the ISSA 2014 Conference*, 2014, pp. 1–6, doi: 10.1109/ISSA.2014.6950509.

[7] G. Tejay, G. Klein, "Organizational Cybersecurity Journal editorial introduction," *Organizational Cybersecurity Journal: Practice Process and People*, vol. 1, no. 1, pp. 1–4, 2021, doi: 10.1108/ocj-09-2021-017.

[8] A. Jain, H. Tailang, H. Goswami, S. Dutta, M. S. Sankhla, R. Kumar, "Social Engineering: Hacking a Human Being through Technology," *IOSR Journal of Computing Engineering*, vol. 18, no. 5, pp. 94–100, 2016, doi: 10.9790/0661-18050594100.

[9] DiamondIT, "The 7 Layers of Cybersecurity," [Online]. Available: https://www.diamondit.pro/7-layers-of-cybersecurity/. [Accessed: Dec. 03, 2022].

[10] Manhattan Tech Support, "The seven layers of IT security," [Online]. Available: https://www.manhattantechsupport.com/. [Accessed: Dec. 02, 2022].

[11] D. Henshel, M. G. Cains, B. Hoffman, T. Kelley, "Trust as a Human Factor in Holistic Cyber Security Risk Assessment," *Procedia Manufacturing*, vol. 3, pp. 1117–1124, 2015, doi: 10.1016/j.promfg.2015.07.186.

[12] A. M. Shabut, K. T. Lwin, M. A. Hossain, "Cyber attacks, countermeasures, and protection schemes - A state of the art survey," in *SKIMA 2016–2016 10th International Conference on Software, Knowledge, Information Management and Applications*, 2017, pp. 37–44, doi: 10.1109/SKIMA.2016.7916194.

[13] R. Hanzu-Pazara, G. Raicu, R. Zagan, "The Impact of Human Behaviour on Cyber Security of the Maritime Systems," *Advanced Engineering Forum*, vol. 34, pp. 267–274, 2019, doi: 10.4028/www.scientific.net/aef.34.267.

[14] R. Ottis, P. Lorents, "Cyberspace: Definition and Implications," in *Proceedings of the 5th International Conference on Information Warfare and Security*, 2010.

[15] J. R. C. Nurse, "Cybercrime and You: How Criminals Attack and the Human Factors That They Seek to Exploit," in *The Oxford Handbook of Cyberpsychology*, A. Attrill-Smith, C. Fullwood, M. Keep, D. J. Kuss, Eds., Oxford: Oxford Library of Psychology, Oxford Academic, 2019, pp. 662–690, doi: 10.1093/oxfordhb/9780198812746.013.35.

[16] D. Craigen, N. Diakun-Thibault, R. Purse, "Defining Cybersecurity," *Technology Innovation Management Review*, vol. 4, no. 10, pp. 13–21, 2014, doi: 10.22215/timreview835.

[17] J. R. C. Nurse, S. Creese, M. Goldsmith, K. Lamberts, "Trustworthy and effective communication of cybersecurity risks: A review," in *2011 1st Workshop on Socio-Technical Aspects in Security and Trust (STAST)*, IEEE, Sep. 2011, pp. 60–68, doi: 10.1109/STAST.2011.6059257.

[18] "Cybercrime cases hit 82pc," [Online]. Available: https://www.thecitizen.co.tz/tanzania/news/business/cybercrime-cases-hit-82pc-2615466. [Accessed: Feb. 04, 2023].

[19] D. Masesa, B. Munyendo, N. Rishad, P. Musuva-Kigen, N. Karumba, *et al.* "Tanzania Cyber Security Report 2016: Achieving Cyber Security Resilience Through Enhancing Visibility and Increasing Awareness," T*anzania Cyber Security Report 2016*, pp. 1–20, 2016. [Online]. Available: http://www.serianu.com/downloads/TanzaniaCyberSecurityReport2016.pdf. [Accessed: Feb. 04, 2023].

[20] AFRIPOL, "African Cyberthreat Assessment Report: Interpol's *Key Insight into Cybercrime in Africa*," [Online]. Available: https://www.interpol.int. [Accessed: Feb. 04, 2023].

[21] TZ-CERT, "TZ-CERT Honeypots Weekly Report," *Dar es Salaam*, [Online]. Available: https://www.tzcert.go.tz/resources-2/reports/. [Accessed: Feb. 04, 2023].

[22] M. Daudi, *Trust in Sharing Resources in Logistics Collaboration*. Düren: Shaker Verlag GmbH, 2019.

[23] M. Laeequddin, B. S. Sahay, V. Sahay, K. A. Waheed, "Trust building in supply chain partners relationship : an integrated conceptual model," *Journal of Management Development*, vol. 31, no. 6, pp. 550–564, 2012, doi: 10.1108/02621711211230858.

[24] M. Lianos, "Social control after Foucault," *Surveillance & Society*, vol. 1, no. 3, pp. 412–430, 2003.

[25] I. Pinyol, J. Sabater-Mir, "Computational trust and reputation models for open multi-agent systems: A review," *Artifical Intelligence Review*, vol. 40, no. 1, pp. 1–25, 2013, doi: 10.1007/s10462-011-9277-z.

[26] J. Riegelsberger, M. A. Sasse, J. D. McCarthy, "The mechanics of trust: A framework for research and design," *International Journal of Human - Computer Studies*, vol. 62, no. 3, pp. 381–422, 2005, doi: 10.1016/j.ijhcs.2005.01.001.

[27] A. Grizard, L. Vercouter, T. Stratulat, G. Muller, "A peer-to-peer normative system to achieve social order," in: *Coordination, Organizations, Institutions, and Norms in Agent Systems II. COIN 2006. Lecture Notes in Computer Science,* vol. 4386*, R. Noriega et al. Eds. Berlin, Heidelberg: Springer*, 2007, doi: 10.1007/978-3-540-74459-7_18.

[28] C. R. Sunstein, "Social Norms and Social Rules," *Coarse-Sandor Institute for Law & Economics Working Papers*, vol. 36, 1996.

[29] L. Rasmussen, S. Jansson, "Simulated Social control for Secure Internet Commerce," in *New Security Paradigms Workshop*, C. Meadows, Ed., ACM, 1996. [Online]. https://www.nspw.org/papers/1996/nspw1996-rasmusson.pdf. [Accessed: Feb. 04, 2023].

[30] A. Capaldo, I. Giannoccaro, "How does trust affect performance in the supply chain? The moderating role of interdependence," *International Journal of Production Economics*, vol. 166, pp. 36–49, 2015, doi: 10.1016/j.ijpe.2015.04.008.

[31] N. P. Nguyen, N. T. Liem, "Inter-Firm Trust Production: Theoretical Perspectives," *International Journal of Business and Management*, vol. 8, no. 7, pp. 46–54, doi: 10.5539/ijbm.v8n7p46.

[32] P. M. Doney, J. P. Cannon, "An Examination of the Nature of Trust in Buyer-Seller Relationships," *Journal of Marketing*, vol. 61, no. April, pp. 35–51, 1997, doi: 10.2307/1251829.

[33] D. M. Rousseau, S. B. Sitkin, R. S. Burt, C. Camerer, "Not so different after all: A cross-discipline view of trust," *Academy of Management Review*, vol. 23, no. 3, pp. 393–404, 1998, doi: 10.5465/AMR.1998.926617.

[34] G. Tejpal, R. K. Garg, A. Sachdeva, "Trust among supply chain partners: A review," *Measuring Business Excellence*, vol. 17, no. 1, pp. 51–71, 2013, doi: 10.1108/13683041311311365.

[35] B. H. Sheppard, D. M. Sherman, "The Grammars of Trust: A Model and General Implications," *The Academy of Management Review*, vol. 23, no. 3, pp. 422–437, 2016, doi:10.2307/259287.

[36] A. F. Salam, L. Iyer, P. Palvia, R. Singh, "Trust in e-commerce," *Communications of the ACM*, vol. 48, no. 2, pp. 72–77, 2005, doi: 10.1145/1042091.1042093.

[37] D. L. Paul, R. R. McDaniel, "A Field Study of the Effect of Interpersonal Trust on Virtual Collaborative Relationship Performance," *MIS Quarterly*, vol. 28, no. 2, pp. 183–227, 2004.

[38] R. J. Lewicki, M. A. Stevenson, R. Lewicki, M. A. Stevenson, "Trust Development in Negotiation: Proposed Actions and a Research Agenda," *Business & Professional Ethics Journal*, vol. 16, no. 1, pp. 99–132, 1997.

[39] J. M. da C. Hernandez, C. C. dos Santos, "Development-based Trust: Proposing and Validating a New Trust Measurement Model for Buyer-Seller Relationships," *Brazilian Administration Review*, vol. 7, no. 2, pp. 172–197, 2010, doi: 10.1590/S1807-76922010000200005.

[40] O. Schilke, G. Wiedenfels, M. Brettel, L. G. Zucker, "Interorganizational trust production contingent on product and performance uncertainty," *Socio-Economic Review*, vol. 1, no. 2, pp. 307–330, 2017, doi: 10.1093/ser/mww003.

[41] E. Jaakkola, "Designing conceptual articles: four approaches," *AMS Review*, vol. 10, no. 1–2, pp. 18–26, 2020, doi: 10.1007/s13162-020-00161-0.

[42] I. H. Bakar, "Social engineering tactics used in mobile money theft in Tanzania," *The University of Dodoma*, 2016. [Online]. Available: http://repository.udom.ac.tz/handle/20.500.12661/1168. [Accessed: Apr. 29, 2023].

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

State-level Cyber Resilience: A Conceptual Framework

Geoffrey A. Hubbard | Bavarian School of Public Policy, Technical University of Munich, Germany, ORCID: 0000-0002-0771-292X

Abstract

There is currently a gap in our academic and practical understanding of the concept of resilience in cyber space at the level of the state, hampering research and policy-making due to the lack of a rigorously constructed, shared terminology. This article contributes to this area by providing a comprehensive capacities-based conceptualisation of state-level cyber resilience. After establishing that cyber resilience is necessary and that it should be developed at the state level, we perform a rigorous exploration of the concept of resilience as it pertains to the different areas involved in state-level cyber resilience. Seeking the most salient characteristics of each one, we identify from the general concept of resilience that it is a non-static process requiring an availability of assets; from state resilience, we identify that resilience capacities are harboured at multiple levels and across actors within the polity; and from cyber resilience, we identify that there is a plethora of different potential damages. Taking all this into consideration, our resulting concept of state-level cyber resilience is the following: the ability of a state, which (a) is made up of multiple layers, to (b) harness a set of key assets in order to (c) confront a particular type of damage to its cyber space, by (d) going through the stages of coping and eventually recovering to its normal state. Having constructed this conceptual framework, this work aids researchers and decision-makers by providing a common terminology and fostering a systematic, multidimensional approach to states' capacity for resilience in cyber-space.

Keywords

cybersecurity, cyberspace, national cyber resilience, critical infrastructure, national security, smart cities

Received: 08.05.2023

Accepted: 03.07.2023

Published: 09.08.2023

Cite this article as:
G.A. Hubbard, "State-level Cyber Resilience: A Conceptual Framework," ACIG, vol. 2, no. 1, 2023, DOI: 10.60097/ACIG/162859

Corresponding author:
Geoffrey A. Hubbard, Bavarian School of Public Policy, Technical University of Munich, Germany;
ORCID: 0000-0002-0771-292X,
E-MAIL: geoffrey.hubbard@tum.de

1. Introduction

We are currently experiencing a paradox: cyber technologies, which were partly created with the goal of making societies more resilient, now harbour threats that jeopardise this objective. Some of the technology that sustains cyber-space emerged from research intended to design communications able to withstand a nuclear attack [1]. Despite this background, cyber-space has inherent characteristics that nevertheless make it vulnerable. Humanity is in the midst of a technological revolution that "challenges all historical experience" [2] and our interconnectedness, for all its benefits, is rendering us ever more vulnerable to "radical... systemic shocks" [3]. As cyber technologies increasingly underpin the functioning of modern societies, states (also known colloquially as 'nations') find themselves in a position of growing dependence. For as long as cyber-space is operating normally, a state can reap its benefits and function as usual. But digital systems are inherently vulnerable, and when compromised, have the potential to severely affect a society's functioning. An example to illustrate the gravity of these vulnerabilities is the 2021 Colonial Pipeline ransomware attack, one of the largest cyber-attacks on American infrastructure and one of the most disruptive digital ransom operations on record.

The Colonial system, running 5,500 miles between Texas and New York, is the largest U.S. gasoline pipeline and transports 2.5 million barrels per day. It is the main source of fuel for the region, carrying nearly half of all fuel consumed on the East Coast [4]. The May 2021 attack compromised IT systems, locking down the victim's computers and demanding payment. A day earlier, the hackers had already stolen confidential industry data, which they threatened to leak if the payment was not made. To contain the attack, the company halted the pipeline's operations along the entire network, and in response to the double extortion attempt, decided to pay the ransom, worth 75 bitcoins. After receiving payment, the hackers granted the company access to its systems, but the recovery process was slow. The shutdown lasted six days, during which uncertainty and panic over fuel supply took hold across the East Coast [5, 6]. Gasoline prices spiked to a six-year high and gas stations continued running out of fuel, even days after the shutdown [7]. In order to best deal with such incidents, a state must have measures in place to improve its resilience. Unfortunately, the main focus in policy so far has been on cyber security (i.e. ensuring systems are fail-safe), and not on cyber resilience (i.e. ensuring systems are safe-to-fail). Furthermore, there remains a paucity of work in the social sciences regarding the intersection of resilience and cyber technologies, with different scholars asserting five years apart that this research topic remains in its infancy [8, 9].

A specific lacuna in our understanding hinders a shift in focus to improve resilience, namely, the fact that an integrated concept of state-level cyber resilience remains inchoate. Under these circumstances, policymakers are left struggling to implement strategies to improve this type of resilience. A consensus has emerged, particularly in Europe and the USA, that cyber-resilience considerations at the level of the state should be included in policy and regulation [10]. To accomplish this objective, further research is needed to better understand the phenomenon. The question that drives this research is therefore the following: how can we conceptualise the term state-level cyber resilience in a firmly grounded and comprehensive manner? This article addresses this gap in research by providing a capacities-based conceptual framework of state-level cyber resilience.

To fulfil its purpose, the article takes the following structure: after this introductory section and a brief note on terminology, we touch upon the theoretical background that supports the key assumptions in conducting this work. Next, we develop a conceptualisation of state-level cyber resilience. To this end, we conduct a conceptual examination of resilience from three points of view, namely the general, cyber, and state perspectives. With this knowledge we then proceed to positing a new concept of statelevel cyber resilience. In closing, we offer a final discussion of the advances made, reflect on future research paths, and share our concluding thoughts.

1.1. Terminology and theoretical assumptions

It is fundamental to have a good understanding of what cyber-space is to be able to study it. Over time there have been various approaches to the concept. Kuehl, for instance, identified fourteen different definitions [11]. For this work, cyberspace is defined as the "fusion of all communication networks, databases, and sources of information into a vast... blanket of electronic interchange" [12]. Importantly, cyber-space is a hybrid construction of physical and virtual layers [13]. As such, the result is a virtual interaction space enabled by new information technologies with physical grounding [14].

We are particularly interested in understanding how this "global synthetic substrate" relates to the functioning of societies [15]. A useful analogy to understand how societies interact with and in cyber-space is to view it as "a globally unfettered exchange space... like an enormous, ... moderately chaotic, annual medieval fair without adequate security from an overlord... and with all the human energy and pathologies possible in shared space" [15]. A plethora of risks fills this domain, varying by cause and including those arising from nefarious

intent, human error or due to environmental circumstances beyond human control. In this work, we collectively refer to the actualisation of these risks as adverse incidents in cyber-space.

This work has two main underlying assumptions, namely, (1) that developing resilience is desirable and necessary, and (2) that the state is an important unit of study related to resilience. Supporting the first assumption, we employ the theoretical background of a risk society. Developed most prominently by Ulrich Beck, he defined this idea as "a systematic way of dealing with hazards and insecurities induced and introduced by modernisation itself" [16]. He stated that dangerous threats to humanity have become an inherent part of industrial life, rather than a manageable by-product. These self-inflicted risks of modernisation are known as manufactured risks. In the same way that human actions are the key cause, humans can also (and must) do much to reduce the threat [17, 18]. Bearing this in mind helps us understand how modern developments that deeply disrupt human life become a double-edged sword. Our stance assumes that states have a resilience deficit regarding the manufactured risks of cyber-space, and that they need to act to improve their capacities for cyber resilience.

For our second assumption, we refer to the evolving role of the state and its ongoing primacy in organised human life. Over the last century, the reach of the state has grown with the development of welfare systems around the world. As such, states have generally taken on greater and more diverse responsibilities than they previously had. In the field of technology, it has been observed that states are highly reliant on the private sector for the development of cyber-capabilities, leading to the question of whether and to what extent the power of the state has been eroded in this area. Nevertheless, however much the power gap narrows between state and non-state actors, there are some key aspects where state power is nevertheless unrivalled – states still exercise the ultimate power of coercion and, unlike private actors, generally have social legitimacy, formal authority, and regulatory capability [14].

Returning to the analogy of cyber-space as a vast medieval fair, we see that despite all the dynamic power these fairs created for private citizens, ultimate control remained in the hands of the state – these markets did not substitute the institutions of feudal authority. We follow Nye's reasoning that cyber-space does not fundamentally challenge the governments of sovereign states, but like medieval markets, it will "coexist and greatly complicate what it means to be a sovereign state" [13]. By this reasoning we assume that states will remain the dominant

actors in cyber-space for the foreseeable future, and hold that state policy can have a substantial effect on cyber resilience capabilities.

2. Conceptual background

Resilience is a complex and multi-faceted phenomenon. As such it has been approached from different angles with variations in its manifestations. In this section we will explore the current understanding of the concept of resilience in different contexts, and develop a comprehensive conceptualisation of state-level cyber resilience. To this end, we first examine the concept of resilience in general, followed by a review of the concepts of state resilience and cyber resilience. Finally, we probe existing contributions in the direction of the joint concept of state-level cyber resilience, integrate the insights from the individual terms, and provide a new conceptualisation.

To lay the foundations for developing a sound conceptualisation of statelevel cyber resilience, it is necessary to start by understanding the background and existing applications of resilience in its general form. Deriving from the Latin resiliere, meaning to bounce back, the word resilience refers to an object's ability to return to its normal state following a disturbance. Over the decades, the concept of resilience has been used and developed to differing degrees in various fields, including materials science, engineering, psychology, ecology, and economics [19]. This co-development across disciplines means that the current understanding of resilience has diffuse roots. In the context of material sciences, resilience can be observed in its most tangible form, and its level is determined by how much stress or force a material can withstand before being permanently altered (e.g. breaking), and how quickly it can return to its previous state once the stress has been removed. The same fundamental logic is applied in the other disciplines, where the object of study is replaced by any other entity that can show a form of resilience construed in a broader sense, ranging from machines to humans and entire ecosystems [20]. As we would expect, the meaning of resilience becomes increasingly nuanced as we move beyond the study of materials into other domains. With its demonstrable permeability across disciplines, resilience is "a polysemous and malleable term" [19]. Despite variations, the fundamental concept outlined above remains the same across applications and there is an overlap of basic features across disciplines [21].

To determine whether an object is showing resilience it is necessary to understand what its normal state is. Therefore, having a clear

definition for the state of normalcy in question is a key consideration for the concept to be of any use. This normalcy can refer, e.g., to physical structure, in the case of materials, or to effective functioning – i.e., the delivery of expected results – in the case of a system. In this work we address resilience specifically as it is applied to human socio-technological systems, and this is the focus we shall take henceforth.

When observing a system, the layers at which resilience manifests itself are multiplied when compared to simpler objects. As a complex whole formed by an interconnected network of elements working together, a system is characterised by interdependence in order to achieve its function. As such, when looking at the resilience of a system, there are two simultaneous approaches [21]. First, it can be understood as the entire system's ability to maintain or resume its functions in spite of a disruption, in other words, how easily the system can run as it should normally when it has been disturbed. Second, it can also be viewed as a sum of multiple instances of resilience of the constituent elements and interconnections within the system. In this sense, one is looking at the maintenance of a sum of sub-functions, that keep the system running as a whole.

When dealing with human systems in particular, such as a company or a state, there is a further consideration, namely that the system and its components are not static at any point, including when disturbed. Human organisations are an example of adaptive systems, where resilience levels and response types vary depending on the characteristics of the component and of the adverse incident [20].

Resilience becomes manifest when an adverse incident occurs and the object of study has to cope with it. Resilience cannot be reduced to a single moment or state of being; rather, it is displayed as a process, involving preparation, detection and response, coping and/or adaptation, and recovery. As such, under normal conditions, resilience cannot be directly observed. Instead, what can be observed is an object's perceived capacity to perform at an acceptable level at each stage of the process of resilience in the event of an adverse incident. In other words, what can be observed in a state of normalcy is an object's perceived potential for resilience.

An important point to bear in mind is that a system's capacity for resilience will usually be influenced by the range of assets it has available. Possessing capacities solely sufficient for functioning in a state of normalcy may be detrimental during an adverse incident. Therefore, there must usually be latent resources that can be called

upon in case of emergency. Additionally, as mentioned above, resilience involves preparation as one of its stages, and this equates to an investment of resources in anticipation of disruptions. Systems resilience, therefore, entails clear costs. These, however, are significantly lower than the potential costs to an organisation that is not resilient. The higher an organisation's capacity for resilience, the lower the unforeseen costs deriving from adverse events. It is for this reason that investing in resilience capacities is described as a form of 'insurance' [20]. Here, an important decision must be made: pay the guaranteed costs of improving resilience, or wait to see the costs of an adverse incident, if it happens at all. Especially when budgets are constrained, it can be tempting to gamble by choosing the unknown costs.

2.1. Cyber resilience

Unlike other areas where resilience has been studied for longer, resilience in relation to the functioning of cyber technologies and its wider implications is relatively recent. Cyber resilience started to gain wide attention from 2012, with a World Economic Forum meeting focusing on the topic [9]. Since then, interest in cyber resilience has grown continuously [22]. The goal of cyber resilience has been described in terms very similar to the general concept of resilience. This concept can have different interpretations, including that of: (a) a purely technical system (e.g. a network) being resilient, or (b) an actor that uses cyber-space being resilient, or (c) a technical system's functions being resilient. Regarding the first approach, Linkov and Kott define it as "the ability of the system to prepare, absorb, recover, and adapt to adverse effects, especially those associated with cyber-attacks" [21]. As for the second approach, and serving as a bridge to the third, Hausken defines it as "the ability of an actor to resist, respond and recover from cyber incidents to ensure the actor's operational continuity" [23]. Continuity is indeed key and, when addressing socio-technological systems, we hold that it is more important to focus on the functions of cyber-space than on the technologies themselves. For Björck et al., cyber resilience refers to the ability to continuously deliver the system's intended outcomes, in spite of disturbances. The intended outcome refers to whatever it is that the technology is meant to deliver to the user. The starting point we take for cyber resilience is thus the use derived from the technology [9]. Bellini and Marrone share this thinking, asserting that cyber resilience seeks to guarantee acceptable levels of service by reducing variability and propagation of disruptions throughout the system [8]. Linkov and Kott also recognise the importance of this when they describe cyber resilience as "a bridge between sustaining operations of the system while ensuring mission execution" [21].

As with general resilience, cyber resilience is usually approached as a subset of (cyber) security [21]. The differences mentioned above are nevertheless evident, in particular when looking at their objectives. As Björck et al. succinctly put it, security seeks to protect IT systems with the intention of making them fail-safe. Resilience, in contrast, seeks to ensure "business delivery", regardless of any adverse events, ensuring the systems are safe-to-fail [9] within a specific time horizon. As Bellini and Marrone assert, developing resilience helps address the "remaining known, but unmitigated, risk as well as enhance the overall ability of the system to respond to unknown or emerging threats" [8].

As mentioned in the Introduction, there has been a multiplication of threats in cyber-space. An interview conducted by Radar Services of 105 security experts concluded that between 2018 and 2025, a 300% increase in cyber-attacks per year is expected [23]. Besides the increase in incidents, though, we must be aware of the enormous variety in the types of disruption they involve, and the real-world damage they can cause. This is important because the resilience response required in each case will be different. To exemplify this diversity, there are three incidents that stand out for their distinct kinds of impact: (1) the 2021 Colonial Pipeline ransomware attack, for instance, led to a six-day shutdown of one of the USA's most important energy infrastructure systems, affecting energy supply and markets; (2) the Stuxnet virus (uncovered in 2010) deployed in Iran altered the functioning of the centrifuges in the Natanz nuclear power plant, delaying the state's nuclear programme with diplomatic and geopolitical implications; and (3) the 2020 United States federal government data breach, sometimes known as the Solar Winds hack, was a major infringement on sensitive information as the result of cyber espionage exploiting software vulnerabilities. By virtue of such examples, it is evident that disruptions in cyber-space have numerous forms with multifarious impacts.

Indeed, instead of only being viewed as a technical affair, cyber resilience is increasingly also approached in a broader sense. In addition to the conventionally considered physical and information elements of cyber systems, Linkov and Kott also assert that human cognitive and social domains are equally interdependent in cyber systems. As such, cyber systems are increasingly viewed as "multi-genre" networks [21]. Further in this regard, Hausken describes cyber resilience as involving "most societal actors" including governments, organisations, individuals, and others, "at most levels of organization" [23]. In this respect, we see a recognition of cyber resilience as a phenomenon requiring multidisciplinary consideration.

2.2. State resilience

Returning to our theoretical background, we have observed states taking on an expanding goal of ensuring citizens' well-being. Whereas historically states prioritised conventional security, now many states seek to ensure broader well-being, including for instance health and economic growth. In its most basic form, the resilience of a state is manifested in the continued preservation of its functioning in spite of adverse incidents that affect its constituent parts (government, population, territory). In the context of governance, one of the predominant definitions of resilience is that of the National Academies of Science (NAS), which matches directly the general definition mentioned above [21]. In this case also, resilience is generally considered as a subset of the broader state security agenda [24]. Current conceptions of state or 'national' resilience are connected with the development of a risk society that serves as our theoretical background. Indeed, over recent years there has been a "resilience creep" into public discourse [25] due to widening security concerns. Certain key events have strengthened this trend, notably the September 11, 2001 attacks in the United States [19], as well as natural disasters and, most recently, the Covid-19 pandemic. In state resilience there are three aspects which are important to highlight: (1) the incorporation of new threats, (2) the multiple layers of the state involved in resilience, and (3) the importance of critical infrastructures. The rationale for selecting these three aspects is as follows: these three aspects constitute a three-part cycle which has to work properly for a modern state to have a robust capacity for resilience. Regarding the first aspect, with the aforementioned expanding role of the state in ensuring citizens' well-being, there are more ways in which the state can fail to deliver. This is exacerbated by the multiplication of manufactured risks as states become more technologically advanced (see theoretical background). This leads us to the second aspect – as was mentioned, states are highly reliant on private actors for their technological prowess. When analysing the complex functioning of a state, it would be unwise to approach it as a monolithic entity. Instead, it is important to view a state as a system with clearly defined parts working together. Finally, this connects with the third element, regarding critical infrastructures. Responsibility for these tends to be shared between private and public actors, and it is through this cooperation that a state can effectively continue to pursue its goals and face a diversity of threats (aspect 1), thus closing the cycle for understanding modern state resilience. In the following paragraphs we will go into more depth regarding each aspect.

With regards to the first aspect covering the multiplication of recognised threats, a researcher we take as our reference point is Fjäder, who emphasises the evolving role of the state as fundamental for

our current understanding of state-level resilience. He observes that states are faced with a broadening variety of threats linked to global interdependence and the rapid pace of change, and the fact that governments have decided to address a growing number of these threats. Indeed, security strategies around the world have increasingly opted for a new paradigm in which they attempt to cover 'all hazards' for 'all of society'. Given that security cannot fully cover such ambitions, developing resilience in tandem becomes key [20].

The expansion of the threats acknowledged by governments and included in security strategies entails a noteworthy shift in thinking. Given that these new risks are often considered unavoidable, in addition to prevention and conventional security efforts, there is an acceptance of managing the threat impact. This entails, in a sense, a change in the social contract between governments and citizens with regards to security, based on the common understanding that there is a partial shift from preventing to coping with adverse incidents [20].

Despite the increasing importance given to it, the concept of state resilience remains ambiguous, given that it must address multiple elements, including appropriate responses to security failures and emergencies, as well as critical infrastructure management. Because of this breadth, implementing resilience at a state level is not entirely straightforward [20]. Resilience is a necessary strategy with the potential of being highly effective, albeit whose implementation requires planning involving allocating responsibilities and defining their scope, designing methods for cooperation and coordination, as well as setting concrete goals.

Concerning the second aspect, the hurdles of implementing resilience at the state level are hardly surprising given the complexity of the state itself. A state can be understood as a system comprising multiple elements working together for a common goal, namely its preservation, and, in modern states, the well-being of its citizens. When viewing the state, we can employ a systems approach in which overall resilience is made up of multiple instances of resilience in its constituent parts. Several researchers have posited approaches along these lines. Darnell, for instance, holds that for a state to have resilience capacities, it must be resilient at multiple levels: individual, [federal] state, local, and federal [25]. Similarly, Walklate et al. propose a typology consisting of the individual, familial, communal, institutional, 'national', regional, and global levels. As such, a state's overall capacity for resilience is multilayered, and is the product of many interconnected "resiliences" [25].

Finally, regarding the third aspect, when we think of the functioning of a contemporary state (and its conditions of normalcy) it is clear that much of it depends on the running of critical infrastructures. These provide services essential to the social and economic well-being of citizens, to government functions and to public security. Among the sectors typically considered as critical infrastructure are: water services, food, energy, communications, transport, health, banking and finance, policing and defence-related assets. In addition, intangible assets such as supply chains are sometimes included [20, 25]. Finally, there is of course the growing importance of cyber-space as a special type of critical infrastructure often supporting the others (see the next section). Across all these different critical infrastructures, their importance lies in the services they deliver. Whereas before there was a greater focus on the protection of physical infrastructure, now there has been a shift to prioritising the infrastructure's function, i.e., the delivery of critical services. This recognition is manifest in a clear policy trend towards protecting critical services across different 'national' security agendas [10]. Indeed, resilience is closely linked to overall state power, generally described as the ability of a state to achieve its goals and influence other actors [26]. As such, Rowland et al. consider resilience to be one of the attributes of a state that is powerful in cyber-space [27].

What is considered a resilient state will vary between states and cultures, but in general terms it would mean a state that is able to cope with adverse incidents in a manner that is locally reasonable, and to adapt and recover to return to a state similar to the one that existed previously. To this end, critical services that enable the functioning of the state need to be maintained throughout the disruption, or at least rapidly reinstated. In the case of developed states, with a larger extent of critical services, the demands on the state are greater, as there are more services which must be guaranteed to operate.

3. Conceptualising state-level cyber resilience

Having covered the concepts of cyber resilience and state resilience, we can now proceed to an informed examination and subsequent construction of the concept of state-level cyber resilience. A couple of similarities between the previous concepts are apparent, specifically that cyber and state resilience both involve complex systems, and the fact that a resilience approach provides a necessary addition to conventional security measures given the latter's limitations. Beyond these similarities, there is a deep convergence. Specifically, state resilience is increasingly reliant on cyber resilience. Indeed, the latter

is becoming so important that contemporary state resilience can no longer be treated independently. The reason for this amalgamation lies in the increased role of cyber-space in the functioning of a state and its critical infrastructures. Digitalisation has meant that the work of governments and the economic activities of a society are on a trend towards reliance on cyber-space. This is not only apparent in terms of communication, but also in the 'smart' integration that is expanding to ever more economic and public sectors, including energy, transport, housing, and education, as well as across businesses and industries [10, 28].

It is surprising that state resilience is often discussed entirely separately from cyber resilience in the literature, given that cyber-space is emerging as the predominant critical service for a state. Fjäder [20], for instance, does not give cyber-space special emphasis amongst critical services; Walklate et al. [25] do not even mention cyber-space in the context of state resilience. Instead, the role of cyber-space is considered as part of broader communications, i.e., as one critical service among many. This, however, is patently changing. By updating our vision and embracing the notion that cyber-space is a ubiquitous substrate supporting ever more aspects of human existence, we realise that this is a special kind of infrastructure and needs to be accorded greater importance. If a state's use of cyber-space is compromised, then the multitude of other critical services that depend on it will also be compromised in a ripple effect. This is acknowledged by Bellini and Marrone, who observe that due to its "tight interdependency and pervasiveness, a fault on the cyber layer provokes a fault in several critical services…" [8]. This risk of "cascading and escalating failures" [29] across many dimensions of society is acknowledged in our theoretical background as one of the main traits of a contemporary risk society [18]. Rather than being one service among many, cyber technologies have become the critical infrastructure of critical infrastructures.

At this point we should highlight that in spite of the convergence of state resilience and cyber resilience, we do not propose a merging of the two concepts. When referring to the state, we must still distinguish between state-level cyber resilience and state resilience. The reason for this is that state resilience remains a broader concept. There are certain types of resilience that are largely independent of cyber-space, referring to ideational and political aspects, such as the resilience of state institutions, or the resilience of a sense of state belonging [27].

3.1. Existing contributions

Before we conceptualise state-level cyber resilience, we shall acknowledge a few existing contributions to this concept. As

we shall see, though valuable, these are altogether rather scant. There has been a growing interest in state-level cyber resilience due, specifically, to the growth of industry 4.0, the pioneering initiatives of some governments (e.g. the United Kingdom), and the proliferation of adverse incidents [28]. In the following section, we will briefly review two academic sources and one governmental policy paper that stand out as relevant. These have been selected from the extremely limited available material because they are the most representative of existing incipient approaches to the concept.

We begin with an article by Tiirmaa-Klaar, providing an overview of the notion of 'national cyber resilience' and what policymakers should consider in order to increase resilience levels. The author recognises that cyber technologies form a networked substrate for communications and all critical economic sectors across the world [10]. Amongst other points, Tiirmaa-Klaar mentions three basic policy areas that need to be covered to build 'national cyber resilience': protecting critical infrastructure, addressing crime in cyber-space, and developing sufficient state-level incident response capabilities. She also asserts that states need comprehensive cyber governance models, as well as ways of assessing and implementing varying policy goals and priorities [10]. Notwithstanding this, the author fails to provide a definition of "national cyber resilience".

Our second reference work is a systematic literature review of cyber resilience and incident response in smart cities by Ahmadi-Assalemi et al. Considering the manner in which the review is conducted, and the fact that parallels can be drawn between a smart city and a wider "cybered" state [15], we deem this work suitable for use as a proxy for our task. The authors conducted a review of primary studies related to the resilience of cyberphysical systems in smart cities and investigated how current cyber-physical systems address digital forensics and incident response [28]. They found that most of the reviewed literature focuses only on subsets of resilience and related concepts in incident response. Specifically, threat 'detection' had a very high incidence rank, along with 'security' and the broad concept of 'attacks'. In contrast, the term 'resilience' ranked low, with some of its constituent stages ranked very low, namely, 'response', and 'recovery' [28]. Furthermore, the review found that many of the papers focused only on particular sectors of a smart city (e.g. infrastructure, mobility), rather than on the cumulative whole [28]. This confirms that there is a dearth of scholarly work on cyber resilience and that the focus has instead been on more conventional security. The article provides only a generic definition for cyber resilience, without expressly connecting it to smart cities or states [28].

Our third and final reference is the United Kingdom's National Cyber Strategy 2022 [30]. This policy paper puts significant emphasis on cyber resilience as a state priority. The UK is one of the states at the forefront of research and policy concerning matters of cyber security. In 2016, the UK set up the National Cyber Security Centre (NCSC) with the task of protecting both the government and society in cyber-space [31]. In spite of actions such as this to improve the state's overall cyber security standing, the policy paper states that there is "growing evidence of gaps in our national resilience", with the number of incidents affecting government, businesses and individuals continuing to rise [30]. With its new strategy the government aims to work towards a vision of cyber-space "as a reliable and resilient place for people and business to flourish" as a fundamental part of building a "more resilient nation" [30]. This apparent level of concern and commitment is significant coming from one of the states considered to be most 'powerful' in cyber-space and highlights the ubiquitous perceived risks states face in cyber-space [32].

Unlike previous iterations, the 2022 strategy includes a definition of cyber resilience from the state's perspective which encompasses systems, organisations, and individuals [30]. Furthermore, the direction taken in this paper shows a maturation from a resilience perspective as it explicitly states the importance of aspects such as having a whole-of-society approach; differentiates between pre- and post-incident measures; stresses the need for collaboration with the private sector, as well as the proactivity of the latter; recognises the importance of (other) critical infrastructures; and highlights the need for government to provide direction and set an example.

With this brief review, we can see how state-level cyber resilience is gaining attention. This growing interest, though, has not yet produced significant theoretical advancements and the concept remains incipient. Indeed, something acknowledged in all three reference works is that further research is needed. The concept is still rudimentary and hardly goes beyond the generic definition of resilience. Without a sound and wellgrounded definition, we run the risk of state-level cyber resilience becoming a vague and misused concept, further clouding attempts for assessment and improvement. With this reasoning in mind, we now proceed to proposing a new, comprehensive concept.

3.2. Conceptual framework

In order to contribute a concept of state-level cyber resilience that can then be operationalised, it must be comprehensive

and concrete. At this point we can identify the elements we need for our conceptualisation. From the general concept of resilience, we know that it requires an availability of assets and an investment of resources, and we have understood that resilience is not static, but is manifested as a process; from state resilience, we know that resilience capacities are harboured at multiple levels and across actors within the polity; and from cyber resilience, we know that there is a wide variety of damage that can be inflicted, which would call for different resilience responses. Taking all this into consideration, our resulting concept is the following:

> state-level cyber resilience: the ability of a state, which (a) is made up of multiple layers, to (b) harness a set of key assets in order to (c) confront a particular type of damage to its cyber space, by (d) going through the stages of withstanding this damage and eventually recovering to its normal state.

The state of normalcy will vary between cases. Nevertheless, in abstract terms, we know that it will be the conditions in which the state finds itself capable of sustaining modern life in its typical day-to-day manner. This entails the provision of critical services, primarily the use of cyber-space, for the state to conduct its core functions. As for the adverse incidents that could occur which require a resilient response from the state, the threats are innumerable. Acknowledging the special trait of resilience as being applicable to unforeseen disturbances, we will consider these adverse incidents as being anything negatively affecting the use of cyber-space within the state, whether caused by humans or nature [9].

In order to operationalise the concept, we must first identify the variables involved. From the concept above, these can be isolated as follows:

- **a.** Layers;
- **b.** Assets;
- **c.** Damage;
- **d.** Stages.

These variables have component indicators that allow for their assessment. A deep exploration of these indicators is beyond the scope of this article, but we will briefly propose a set to illustrate the concept's operationalisation. These indicators have been selected with the intention of being comprehensive with respect to the key elements of each type of variable, whilst being succinct and

thus making assessment straightforward. In doing this we have heeded the recommendation that resilience metrics should be (1) broad enough to be used in diverse cases and (2) precise enough to measure specific system components [33]. Cyber resilience is "flexible by nature" [33] and as such, we reason that it is an adequate approach to provide the evaluator with a degree of autonomy within the framework.

Layers

Governments may be the directing actors within a state, but improving state cyber resilience requires multiple actors working together. Hausken, for instance, names eight state layers involved in cyber resilience [23]. We consider this selection inconsistent with our unit of analysis and therefore propose our own set of four layers where resilience is manifested, consisting of the government, as the directing and coalescing actor; private companies, as the main organised entities performing economic activity; communities, as the main organised entities performing non-economic activity, and the individual, as the smallest and most numerous unit within a state.

Assets

As discussed above, resilience has a cost, requiring an investment of resources in anticipation of disruptions. When this investment is effective, it means that the state in question can deploy or activate a number of assets to support its resilience response. From the existing literature, we will base our approach on the set of key resilience assets posited by Bellini and Marrone, consisting of human capital, involving the level of skills and preparedness of the people; technology, which includes the cyber technologies involved in the incident; organisation, referring to how well the states' layers can cooperate; and finance, referring to the capital at the state's disposal for confronting an adverse incident [8].

Damage

When it comes to distinguishing between types of damage, we suggest employing the CIA triad of cybersecurity, a common classification for the kind of damage inflicted in cyber-space. This acronym stands for the damage that can be suffered with regards to Confidentiality, Integrity, and Availability of data or systems. The impact of each type of damage would have to be assessed in relation to the state of normalcy of the state being studied, at a particular time [9, 24].

Table 1. State-level cyber resilience variables and their proposed component indicators.

Layers	Assets	Damage	Stages
Government	Human capital	Confidentiality	Preparation
Companies	Technology	Integrity	Response
Communities	Organisation	Availability	Recovery
Individuals	Finance		

Stages

As discussed earlier, when resilience is put into practice, it manifests itself as a process before, during and after an adverse incident. A common typology of stages is that employed by Bellini and Marrone [8], namely: prepare, withstand or absorb, recover and adapt. Although we find this typology to be insightful, we prefer a slightly condensed version consisting of three stages: Preparation, Response, and Recovery. Here we consider the Response stage to include both 'absorbing' a shock, as well as 'adapting' to it for its duration.

Much of the challenge in addressing the resilience this article has tackled stems from the fact that cyber resilience is typically understood pertaining to individual parts of the state system, and had not yet been conceptualised at the system level, incorporating the different constituent elements. This article explores and analyses the most important aspects of resilience, and subsequently distils them into an integrated and concise concept.

This framework will aid scholars and policymakers in identifying areas of strength and weakness in states' resilience, and the insights it provides will inform strategic decision-making and resource-allocation. In particular, it helps to avoid the potential quagmire of addressing resilience in a siloed manner. It simplifies approaching the issue by extracting the four most salient variables and describing how they interrelate to form a single concept. We provide a way of operationalising the concept by means of a set of indicators which serve as suggested guideposts for a comprehensive step-bystep assessment. With this four-pronged conceptual framework, the different elements of resilience can be approached simultaneously, allowing for research and policymaking that takes into consideration the full picture of state-level cyber resilience. This approach does not point at specific solutions. As Shimizu and Clark point out, "linear

and fixed decision-making approaches are of limited value" due to complexity and uncertainty [18]. Rather, our framework leaves the interested parties with the necessary flexibility and freedom to create their own strategies for improvement based on the specific insight that the assessment provides.

To illustrate the utility of this conceptual framework, we will return to the infamous case of the Colonial Pipeline ransomware attack. In the following table, we present the types of questions that could arise for each variable in this scenario.

Table 2. Illustrative application of the concept to a scenario based on the Colonial Pipeline attack

Variable	Indicators	Example questions for the case
Layers	Government Companies Communities Individuals	What are their responsibilities in this scenario? What are their strengths and weaknesses? How can these layers prepare to increase their capacity for resilience?
Assets	Human capital Technology Organisation Finance	How can the different layers of the state in question harness these assets in such a scenario? Do the layers have the necessary skills? What is the condition of the relevant technology? Are there mechanisms in place for effective cooperation within and across the relevant layers? What is the financial landscape and how would it respond to such a scenario?
Damage	Confidentiality Integrity Availability	What types of damage will the state suffer? Which of these would be most harmful? Which one is most likely and what measures are in place to deal with such damage?
Stages	Preparation Response Recovery	Given the previous questions and answers, what is the assessment of the state's overall preparation? Based on this, what is the perceived competence for a response and recovery to such an incident?

4. Conclusions

This work has taken on an ambitious challenge. In an increasingly important research field that is nevertheless in its infancy, we have proposed a comprehensive conceptual framework of state-level cyber resilience. To accomplish this, we have relied on an intensive cross-pollination of ideas and information provided by other scholars in related research areas. We do not claim to have achieved a definitive concept of state-level cyber resilience; rather, the accomplishments of this work are to aid researchers and policy-makers by providing a common terminology, fostering a systematic and multidimensional approach to states' capacity for resilience in cyber-space, and supplying a springboard for academic debate and further research.

A fascinating ensuing line of research would be to examine how the level of complexity of states aids or hinders their cyber resilience. Complexity has been observed to both strengthen and weaken resilience in systems [21], and states are no exception. Investigating the nature of this simultaneous scope for benefit and detriment would contribute greatly to this field's solidity.

Resilience as a strategy is not a panacea for state security challenges relating to cyber-space and beyond. It nevertheless provides a unique advantage by addressing unpreventable security challenges, whilst also being cheaper in the long term than conventional security. Total resilience cannot be guaranteed, even when adequate strategies are implemented, but a comprehensive understanding of state-level cyber resilience would nonetheless provide much-needed insight so that states can improve their resilience potential. The conceptual framework provided in this work is a step in this direction.

References

[1] J. Ryan, *A history of the Internet and the digital future*. London: Reaktion Books, 2010.

[2] H. Kissinger, *World order*. New York: Penguin Press, 2014.

[3] World Economic Forum, The Global Risks Report 2018, 2018. [Online]. Available: http://www3.weforum. org/docs/WEF_GRR18_Report.pdf. [Accessed: May 4, 2023].

[4] C. Bing, S. Kelly, "Cyber attack shuts down U.S. fuel pipeline 'jugular,' Biden briefed," [Online]. Available: https://www.reuters.com/technology/colonial-pipeline-halts-all-pipeline-operations-after-cybersecurity-attack-2021-05-08/. [Accessed: Dec. 1, 2021].

[5] U.S. Government Accountability Office, "Colonial Pipeline Cyberattack Highlights Need for Better Federal and Private-Sector Preparedness (infographic)," U.S. GAO, May 18, 2021. [Online]. Available: https://www.gao.gov/blog/colonial-pipeline-cyberattack-highlights-need-better-federal-and-private-sector-preparedness-infographic. [Accessed: May 4, 2023].

[6] U.S. Department of Justice, "Department of Justice Seizes $2.3 Million in Cryptocurrency Paid to the Ransomware Extortionists Darkside," June 7, 2021. [Online]. Available: https://www.justice.gov/opa/pr/department-justice-seizes-23-million-cryptocurrency-paid-ransomware-extortionists-darkside. [Accessed: May 4, 2023].

[7] C. Thorbecke, "Gas hits highest price in 6 years, fuel outages persist despite Colonial Pipeline restart," [Online]. Available: https://abcnews.go.com/us/gas-hits-highest-price-years-fuel-outagespersist/story?id=77735010. [Accessed: May 5, 2023].

[8] E. Bellini, S. Marrone, "Towards a novel conceptualization of Cyber Resilience," *2020 IEEE World Congress on Services (SERVICES)*, pp. 189–196, 2020, doi: 10.1109/SERVICES48979.2020.00048.

[9] F. Björck, M. Henkel, J. Stirna, J. Zdravkovic, "Cyber Resilience – Fundamentals for a Definition," in *New Contributions in Information Systems and Technologies*, vol. 353, A. Rocha, A. M. Correia, S. Costanzo, and L. P. Reis, Eds. Cham: Springer International Publishing, pp. 311–316, 2015, doi: 10.1007/978-3-319-16486-1_31.

[10] H. Tiirmaa-Klaar, "Building national cyber resilience and protecting critical information infrastructure," *Journal of Cyber Policy*, vol. 1, no. 1, pp. 94–106, 2016, doi: 10.1080/23738871.2016.1165716.

[11] D. Kuehl, "From Cyberspace to Cyberpower: Defining the Problem," in *Cyberpower and National Security*, F. D. Kramer, S. H. Starr, L. K. Wentz, Eds. 1st ed.Washington, DC: National Defense University Press, 2009, pp. 24–42.

[12] M. Dunn Cavelty, "Cyber-Security," in *Contemporary Security Studies*, A. Collins, Ed., Oxford: Oxford University Press, 2016, pp. 400–416.

[13] J. S. Nye, *The future of power*, 1st ed. New York: Public Affairs, 2011.

[14] N. Choucri, "Emerging Trends in Cyberspace: Dimensions & Dilemmas," in Cyberspace: Malevolent Actors, Criminal Opportunities and Strategic Competition, 2012, pp. 1–19. [Online]. Available: https://nchoucri.mit.edu/sites/default/files/documents/[Choucri]%202012%20Emerging%20Trends%20in%20CyberspaceDimensions%20%26%20Dilemmas.pdf. [Accessed: May 5, 2023].

[15] C. Demchak, "Cybered Conflict, Cyber Power, and Security Resilience as Strategy," in *Cyberspace and national security: threats, opportunities, and power in a virtual world*, D. S. Reveron, Ed., Washington, DC: Georgetown University Press, 2012, pp. 121–136.

[16] U. Beck, "Risk society: towards a new modernity," in *Theory, culture & society*. London, Newbury Park, New Delhi: Sage Publications, 1992.

[17] U. Beck, *World at risk*. Cambridge: Polity Press, 2009.

[18] M. Shimizu, A. L. Clark, *Nexus of Resilience and Public Policy in a Modern Risk Society*. Singapore: Springer Singapore, 2019. doi: 10.1007/978-981-10-7362-5.

[19] T. Prior, J. Hagmann, "Measuring resilience: methodological and political challenges of a trend security concept," *Journal of Risk Research*, vol. 17, no. 3, pp. 281–298, 2014, doi: 10.1080/13669877.2013.808686.

[20] C. Fjäder, "The nation-state, national security and resilience in the age of globalisation," Resilience, vol. 2, no. 2, pp. 114–129, 2014, doi: 10.1080/21693293.2014.914771.

[21] I. Linkov, A. Kott, "Fundamental Concepts of Cyber Resilience: Introduction and Overview," in *Cyber Resilience of Systems and Networks*, A. Kott, I. Linkov, Eds. Cham: Springer International Publishing, 2019, pp. 1–25. doi: 10.1007/978-3-319-77492-3_1.

[22] D. A. Sepúlveda Estay, R. Sahay, M. B. Barfod, C. D. Jensen, "A systematic review of cyber-resilience assessment frameworks," *Computers & Security*, vol. 97, 2020, doi: 10.1016/j.cose.2020.101996.

[23] K. Hausken, "Cyber resilience in firms, organizations and societies," *Internet of Things*, vol. 11, 2020, doi: 10.1016/j.iot.2020.100204.

[24] E. G. Carayannis, E. Grigoroudis, S. S. Rehman, N. Samarakoon, "Ambidextrous Cybersecurity: The Seven Pillars (7Ps) of Cyber Resilience," IEEE Trans. Eng. Manage., vol. 68, no. 1, pp. 223–234, 2021, doi: 10.1109/TEM.2019.2909909.

[25] S. Walklate, R. McGarry, G. Mythen, "Searching for Resilience: A Conceptual Excavation," *Armed Forces & Society*, vol. 40, no. 3, pp. 408–427, 2014, doi: 10.1177/0095327X12465419.

[26] A. F. K. Organski, *World politics*, 2nd ed. New York: Alfred A. Knopf, 1968.

[27] J. Rowland, M. Rice, S. Shenoi, "The anatomy of a cyber power," *International Journal of Critical Infrastructure Protection*, vol. 7, no. 1, pp. 3–11, 2014, doi: 10.1016/j.ijcip.2014.01.001.

[28] G. Ahmadi-Assalemi, H. Al-Khateeb, G. Epiphaniou, C. Maple, "Cyber Resilience and Incident Response in Smart Cities: A Systematic Literature Review," *Smart Cities*, vol. 3, no. 3, pp. 894–927, 2020, doi: 10.3390/ smartcities3030046.

[29] A. Vespignani, "The fragility of interdependency," *Nature*, vol. 464, no. 7291, pp. 984–985, 2010, doi: 10.1038/464984a.

[30] H.M. Government, "National Cyber Strategy 2022," [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1040805/National_Cyber_Strategy_-_FINAL_VERSION.pdf. [Accessed: May 5, 2023].

[31] M. Flournoy, M. Sulmeyer, "Battlefield Internet," *Foreign Affairs*, vol. 97, no. 5, pp. 40–46, 2018.

[32] J. Voo, I. Hemani, S. Jones, D. Winnona, D. Cassidy, et al., "National Cyber Power Index 2020," [Online]. Available: https://www.belfercenter.org/publication/national-cyber-power-index-2020. [Accessed: Dec. 10, 2021].

[33] I. Linkov, D. A. Eisenberg, K. Plourde, T. P. Seager, J. Allen, A. Kott, "Resilience metrics for cyber systems," *Environ Syst Decis*, vol. 33, no. 4, pp. 471–476, 2013, doi: 10.1007/s10669-013-9485-y.

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Protection of the EU's Critical Infrastructures: Results and Challenges

Robert Mikac | Faculty of Political Science, University of Zagreb, Croatia,
ORCID: 0000-0003-4568-6299

Abstract

At the end of 2022 and the beginning of 2023, the EU adopted several new legislative acts aimed at improving the resilience and protection of network and information systems and critical entities across the Union. The objective of this research is to list these acts, show their mutual connections and focus specifically on analysing the potential weaknesses of two legislative acts, namely: the NIS2 Directive and the CER Directive. The NIS2 Directive is a significant piece of legislation that aims to improve the cybersecurity of the European Union, while the CER Directive is a crucial piece of legislation that aims to improve the physical security of critical entities in the Union. These two documents are applied in parallel and contain many mutual references, which means that weaknesses in one document may have significant consequences for the implementation of the other. Using standard desktop analysis of primary and secondary sources, this paper reviews results and challenges in the protection of the EU's critical infrastructures by primarily focusing on these two documents. The research identifies and explains certain weaknesses, concluding with suggestions for possible solutions.

Keywords
critical infrastructures, EU legislative framework, NIS2 Directive, CER Directive, results and challenges

Received: 30.08.2023

Accepted: 22.12.2023

Published: 31.12.2023

Cite this article as:
R. Mikac "Protection of the EU's Critical Infrastructures: Results and Challenges," ACIG, vol. 2, no. 1, 2023, DOI: 10.60097/ACIG/162868.

Corresponding author:
Robert Mikac, Faculty of Political Science, University of Zagreb, Croatia; ORCID: 0000-0003-4568-6299; E-MAIL: robert.mikac@fpzg.hr

1. Introduction

"Our daily lives depend on a wide variety of services – such as energy, transport, and finance, as well as health. These rely on both physical and digital infrastructure" [1, p. 2]. Physical infrastructure enables us to work, travel and benefit from essential public services such as hospitals, transport and energy supplies. In contrast, digital infrastructure facilitates a multitude of new, extremely exciting and unpredictably dynamic jobs and processes, virtual realities and artificial intelligence, freedom of speech and possibilities of action. On one hand, all this progress makes the world a pleasant place to live, while, on the other, it raises many questions, many of which contain a security component. What happens in cyberspace has numerous implications for the physical world since "cyber" has become part of the physical reality all around us. Therefore, efficient functioning of physical and digital infrastructure has become one of the key areas of security for individuals, many economies, states, companies of all profiles and large multinational organisations such as the European Union.

The protection of critical physical and digital infrastructure is one of the key areas of national security for many countries around the world and one of the key security priorities of the European Union. The *2020 EU Security Union Strategy* emphasises four strategic priorities for the Security Union, namely: "(i) a future proof security environment, (ii) tackling evolving threats, (iii) protecting Europeans from terrorism and organised crime, (iv) a strong European security ecosystem" [1, p. 6]. The first strategic priority discusses achievements and challenges related to critical physical and digital infrastructure and states "if these infrastructures are not sufficiently protected and resilient, attacks can cause huge disruption – whether physical or digital – both in individual Member States and potentially across the entire EU" [1, p. 6]. Due to the discussion that follows, it is important to point out that the Strategy asserts that "the EU's existing framework for protection and resilience of critical infrastructures has not kept pace with evolving risks" [1, p. 6] with respect to which two directives are considered under the existing framework: *Council Directive 2008/114/EC of 8 December 2008 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection,* and *Directive (EU) 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.*

The security risks are very diverse. Many authors believe that the digital revolution is transforming every aspect of our lives, both

creating enormous opportunities but also increasing exposure to threats [1; 6; 7; 8; 9; 14; 17]. Ulrik Franke et al. state that "with poor cyber security, society is vulnerable, both to accidents and to attacks" [3, p. 116]. Alok Mishra and associates believe that "cyber threats have risen as a result of the growing trend of digitalisation and excessive reliance on the digital world" [4, p. 1]. Sam Maesschalck and associates argue that Industrial Control Systems were not designed with internet connectivity in mind, and often lack basic security features, making them vulnerable to cyberattacks [5]. In addition to the possibility of technical failures, Stefan Varga and associates recognise people as the main sources of security risks to critical infrastructures.

> "People can either (i) take inadvertent unintentional actions, without having a malicious or harmful intent, e.g., by doing mistakes, errors and omissions, (ii) fail to take action in a given situation, where actions otherwise would have prevented an undesired outcome, or (iii) act deliberately with the intent to do harm, e.g., by acts of fraud, sabotage, theft and vandalism" [10, p. 2].

Johan David Michels and Ian Walden identified the risks to critical infrastructure from under-investment in cybersecurity measures and insufficient information sharing [11]. Although critical physical and digital infrastructure areas have become extremely connected over the last 10 years, the focus of policy makers, academics and practitioners has mostly been directed towards the latter area, as well as cyberspace. This is understandable because, as Tomasz Aleksandrowicz says, "cyberspace is now the basis for the functioning of a state's critical infrastructure" [12], both digital and physical.

Although these two areas are inextricably linked, to the best of our knowledge, in the academic world the primary focus is on considering the effectiveness of cyber protection of critical infrastructures and comparing different legal instruments related to it. Dimitra Markopoulou, Vagelis Papakonstantinou, and Paul de Hert discuss the new EU cybersecurity framework, a new Regulation on ENISA (the EU Cybersecurity Act), and the relationship between the NIS1 Directive and the EU's General Data Protection Regulation (GDPR) [13]. The interplay between the NIS1 Directive and the GDPR in a cybersecurity threat landscape is a topic dealt with by Mark D. Cole and Sandra Schmitz-Berndt [14]. Further on the same topic, Sandra Schmitz-Berndt and Stefan Schiffner analyse reporting obligations, as well as certain limitations and differences, with respect to the NIS1 Directive and GDPR [15]. Sandra Schmitz-Berndt then discusses mandatory cybersecurity

incident reporting under the NIS2 Directive [16], and the reporting threshold for a cybersecurity incident under the NIS1 Directive and NIS2 Directive [17]. Comparison of the NIS1 Directive and NIS2 Directive was carried out through several different studies [18; 19; 20]. Finally, it is worth highlighting the comparison of the NIS2 Directive Proposal with the development of Italian and German cybersecurity laws [21]. None of these studies included a comparison of legal instruments related to the protection of critical physical infrastructures.

Returning again to the *2020 EU Security Union Strategy*, which emphasises that "the EU's existing framework for protection and resilience of critical infrastructures has not kept pace with evolving risks", it is further stated that "the legislative framework needs to address this increased interconnectedness and interdependency, with robust critical infrastructure protection and resilience measures, both cyber and physical" [1, p. 6].

> "At the same time, Member States have exercised their margin of discretion by implementing existing legislation in different ways. The resulting fragmentation can undermine the internal market and make cross-border coordination more difficult – most obviously in border regions. Operators providing essential services in different Member States have to comply with different reporting regimes" [1, p. 6].

Therefore, it was particularly emphasised that the European Commission is looking for new legal frameworks for both physical and digital infrastructures [1, pp. 6–7]. After several years of work, at the end of 2022, three legislative acts were published in the same Official Journal of the European Union, namely, two directives and one regulation: (i) *Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS2 Directive)* [22] (hereinafter: NIS2 Directive); (ii) *Directive (EU) 2022/2557 of the European Parliament and of the Council of 14 December 2022 on the resilience of criticzal entities and repealing Council Directive 2008/114/EC* [23] (hereinafter: CER Directive); and (iii) *Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014, (EU) No 909/2014 and (EU) 2016/1011* [24] (hereinafter: DORA).

The aim of this paper is analysing two complementary documents – the NIS2 Directive and CER Directive – because the first document is

focused on improving the cybersecurity of network and information systems across the European Union, and the second on strengthening the resilience and protection of critical entities across the Union.[1] The rationale for such a study design is twofold. First, in normative terms, the NIS2 Directive represents the central document for cybersecurity in the EU in connection with strengthening the resilience and protection of network and information systems, and the CER Directive in the area of strengthening the resilience and protection of critical entities. Second, on an operational and implementation level, network and information systems and critical entities represent one of the key bloodstreams of the Union, Member States, numerous organisations, all economies and a growing number of citizens. That is why it is essential to analyse the strengths and weaknesses of these two documents, which is the narrower purpose of this research. It is clear that both directives have resulted in numerous changes and improvements in the existing normative framework and will lead to better vertical and horizontal operational solutions. However, there is always room for additional work on the quality of legislative acts, which brings us to the question posed by this research: What are the weaknesses of these two directives?

With respect to structure, the *Introduction* is followed by a second section called *From Council Directive 2008/114/EC to the CER Directive*, which will analyse the two above-mentioned directives. The next section, *From the NIS1 Directive to the NIS2 Directive*, will provide an analysis of the two directives in question, followed by a section titled *Discussion*, which will connect the research results from the perspective of the research question and discuss the findings and their implications within the broader context of protecting critical EU infrastructures. *Conclusion* will summarise the analysis and all segments of the research, and provide final comments as well as the significance of the findings.

2. From Council Directive 2008/114/EC to the CER Directive

Council Directive 2008/114/EC of 8 December 2008 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection [25] (hereinafter: Council Directive 2008/114/EC) was the first step in a multi-step process to identify and designate critical European infrastructures and assess the need to improve their protection. As such, this Directive was focused on the energy and transport sectors [25: recital 5]. This Directive set out a procedure for the identification and designation

1 — An analysis of DORA is beyond the scope of this research, so here we only outline its key features and links to the NIS2 Directive. The objective of DORA is to strengthen information security and cybersecurity in the financial sector to maintain operational resilience in case of serious operative disruptions. It refers to entities that operate in the financial sector and third parties that provide services related to information and communication technologies. The NIS2 Directive will also apply to financial institutions (key sector – banking), whereas DORA is lex specialis. The NIS2 Directive aims to secure the resilience of essential and important entities in terms of cybersecurity, and DORA is intended to strengthen the security of financial entities.

of critical European infrastructure and a common approach to assessing the need to improve the protection of such infrastructure to contribute to the protection of people [25: article 1]. It was pointed out that there are a certain number of critical infrastructures at the level of the Union (at the time of the publication of the Directive, the term "Community" was used), the disruption or destruction of which would have significant cross-border impacts. This may include transboundary cross-sector effects resulting from interdependencies between interconnected infrastructures. Such infrastructure should be identified and designated through a common procedure [25: recital 7]. It was additionally emphasised that this Directive complements existing sectoral measures at the level of the Community and Member States [25: recital 10], and that the primary and ultimate responsibility for protecting critical European infrastructures rests with the Member States and the owners/operators of such infrastructures [26: recital 6].

The Directive was relatively short, with only a few articles. The first article sets out the purpose of the Directive, and the second provides definitions. The third article defines the criteria for identifying critical infrastructures, while the fourth article sets out the methods for designating critical infrastructures. The fifth article describes the purpose of the Operator security plan and who is responsible for creating said plan, while the sixth describes the function of Security Liaison Officers. The seventh article refers to reporting of generic data by Member States to the Commission on a summary basis on the types of risks, threats and vulnerabilities encountered in the energy and transport sectors in which critical European infrastructure has been designated. The eighth article states that the European Commission shall support, through the relevant Member State authority, the owners/operators of designated critical European infrastructures by providing access to available best practices and methodologies, as well as training and the exchange of information on new technical developments related to critical infrastructure protection. The ninth article describes the requirements for the protection of sensitive information relating to critical infrastructure protection. The tenth article requires Member States to designate contact points for the protection of critical infrastructure [25].

As the threat landscape has evolved over time, with the emergence of new threats and the increasing interconnectedness of physical and cyber domains, questions have arisen regarding the continued relevance of the current Directive and the need for an update. The European Commission has prepared a comprehensive evaluation study on the scope of the Directive.[2] According to the 2019 *Evaluation*

2 —— This report has been prepared by EY and RAND Europe for the European Commission's Directorate-General for Migration and Home Affairs (DG HOME). The author of this text has been interviewed several times by companies that are preparing a report on the achievements of the current Directive, its weaknesses, its implementation in national legislation, public-private partnerships in the protection of critical infrastructure, and ideas for creating a new Directive.

study of Council Directive 2008/114 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection 10 years after entering into force, "the Directive appears today to have partial to limited relevance, notably in view of recent technological, economic, social, policy/political and environmental developments and current challenges" [26, p. 1]. The Member States adopted a variety of approaches in transposing the Directive in their national legislation [27, p. 14]. Member States have different starting points and approaches towards the identification of potential critical European infrastructure [27, p. 17], while the process of designating critical European infrastructure tends to be less formalised [27, p. 19]. Regarding the Operator security plan, "each Member State adopted this provision using their own interpretations of what needed to be done; this has led to the adoption of different criteria for use in assessing risks for each Member State" [27, p. 21]. The next challenge was the criteria for determining the Security Liaison Officer. Member States applied the requirements that the Security Liaison Office should satisfy very differently (in terms of role, key responsibilities, clearance, etc.) [27, pp. 21–22]. The same was true of very similar claims for national contact points for the protection of critical infrastructure [27, p. 23]. Regarding the reporting of Member States to the Commission, the procedure was established; however, it lacked sufficiently high-quality use of information collected by the Commission, which "has not systematically provided feedback on these reports, nor has it worked to synthesise the situational pictures at the MS level in order to create a pan-EU assessment of critical infrastructure vulnerability" [27, p. 22]. The final assessment is how the Directive "appears to be broadly consistent with relevant sectoral legislation. However, its coherence is limited by the existence of several overlaps with other pieces of legislation and policy documents." Additionally, "the Directive has been partially effective in achieving its stated objectives." Also, the evaluation found that the Directive generated some EU added value [26, pp. 2–6]. For all these reasons, it was necessary to adopt a new and updated directive.

The CER Directive was adopted to eliminate weaknesses observed during the evaluation of Directive 2008/114/EC, which was carried out in 2019 and found that,

> "due to the increasingly interconnected and cross-border nature of operations using critical infrastructure, protective measures relating to individual assets alone are insufficient to prevent all disruptions from taking place. Therefore, it is necessary to shift the approach towards ensuring that risks are better accounted for, that the role and duties of critical

entities as providers of services essential to the functioning of the internal market are better defined and coherent, and that Union rules are adopted to enhance the resilience of critical entities" [23, recital 2].

The CER Directive has repeatedly upgraded the scope and reinforced the resilience and protection of critical infrastructures through various measures and activities.

It is necessary to highlight the key differences between Council Directive 2008/114/EC and the CER Directive:

- **Scope and Coverage:** Council Directive 2008/114/EC focuses exclusively on the energy and transport sectors, while the CER Directive expands the scope to include all critical entities that provide essential services in 11 sectors [23, annex 1].
- **Risk Assessment Approach:** Council Directive 2008/114/EC emphasises a top-down, risk-based approach to critical infrastructure protection, while the CER Directive encourages a more holistic, multi-dimensional approach that considers both physical and cyber threats [23, article 5 and 12].
- **Security Plan Requirements:** The security plan requirements in Council Directive 2008/114/EC were limited to the energy and transport sectors, while the CER Directive mandates the development of security plans for all critical entities, regardless of sector [23, article 13].
- **Incident Response and Recovery Mechanisms:** Council Directive 2008/114/EC provides limited guidance on incident response and recovery, while the CER Directive emphasises the need for robust incident response and recovery plans to ensure continuity of critical services [23, article 15].
- **Information Sharing and Cooperation:** Council Directive 2008/114/EC encourages information sharing between Member States and the Commission, while the CER Directive strengthens this requirement by establishing a centralised information-sharing platform (a Critical Entities Resilience Group is hereby established) and promoting the exchange of best practices and lessons learned [23, article 19].
- **Review and Update Mechanism:** Council Directive 2008/114/EC does not explicitly provide for a regular review and update

> process, while the CER Directive mandates the establishment of a review mechanism to ensure that the Directive remains relevant and effective considering changing threat landscapes and technological advancements [23, article 20].

Overall, the CER Directive represents a significant update to Council Directive 2008/114/EC, reflecting the evolving nature of critical infrastructure threats and the growing importance of a holistic approach to critical infrastructure protection. The CER Directive is a crucial piece of legislation that aims to improve the physical security of critical entities in the European Union, where critical entities represent providers of essential services, and play an indispensable role in the maintenance of vital societal functions or economic activities in the internal market in an increasingly interdependent Union economy. As in the example of the NIS2 Directive, the CER Directive aims at better regulation and alignment of differences between the entities involved in the provision of essential services, which are increasingly subject to diverging security requirements imposed under national law. Therefore, the new Directive seeks to lay down harmonised minimum rules to ensure the provision of essential services in the internal market, to enhance the resilience of critical entities and to improve cross-border cooperation between competent authorities. This Act also recognises other challenges relevant to the regulation of this area and considerable effort has gone into their resolution and normative improvements.

3. From the NIS1 Directive to the NIS2 Directive

Directive (EU) 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 [28] (hereinafter: NIS1 Directive) was the first horizontal legal instrument undertaken at an EU level for the protection of network and information systems across the Union [3; 11; 12; 13; 14; 16; 17; 18; 20; 29]. The Directive aimed to achieve the following:

> "(a) [to lay] down obligations for all Member States to adopt a national strategy on the security of network and information systems; (b) [to create] a Cooperation Group in order to support and facilitate strategic cooperation and the exchange of information among Member States and to develop trust and confidence amongst them; (c) [to create] a computer security incident response teams network ('CSIRTs network') in order to contribute to the development

> of trust and confidence between Member States and to promote swift and effective operational cooperation; (d) [to establish] security and notification requirements for operators of essential services and for digital service providers; (e) [to lay] down obligations for Member States to designate national competent authorities, single points of contact and CSIRTs with tasks related to the security of network and information systems" [28, article 1].

The purpose of the NIS1 Directive was to enhance the security of network and information systems (NIS) across the European Union. The Directive sets out minimum cybersecurity requirements for operators of essential services (OES), digital service providers (DSP), and Member States. The Directive's key objectives were to: (a) reduce the risk of cyberattacks on NIS; (b) improve the incident response capabilities of OES and DSP; (c) strengthen cooperation and information exchange between Member States and the European Commission; (d) foster cross-border cooperation in investigating and prosecuting cyber-crime. The Directive's main requirements for OES and DSP were: (a) identifying and classifying NIS; (b) implementing appropriate security measures to protect NIS; (c) reporting security incidents promptly to the relevant authorities; (d) conducting regular security assessments; (e) developing and implementing incident response plans; (f) providing regular updates on their cybersecurity measures. The Directive also requires Member States to: (a) establish national NIS authorities to oversee the implementation of the Directive; (b) develop and implement national NIS strategies; (c) foster public-private cooperation on cybersecurity; (d) support OES and DSP in implementing the Directive's requirements; (e) investigate and prosecute cyber-crime effectively. The Directive additionally established a framework for mutual assistance between Member States in the event of a major cybersecurity incident. Finally, the Directive required the European Commission to: (a) monitor the implementation of the Directive and provide guidance to Member States; (b) support the development of cybersecurity standards and best practices; (c) promote international cooperation on cybersecurity [28]. The NIS1 Directive was a significant step towards improving the security of NIS in the Union. It is expected to contribute to the resilience of the Union's critical infrastructure and the protection of its citizens.

The obligations under the NIS1 Directive can be broadly divided into two categories: safeguarding obligations, which require organisations to put in place "appropriate and proportionate" security measures, and information obligations, which require the sharing

or disclosure of information [11, p. 16]. While the NIS1 Directive increased the OES, DSP, and Member States' cybersecurity capabilities, its implementation proved difficult. Member States adopted different approaches, resulting in fragmentation at different levels across the internal market [2, 5]. "Different actors understand cybersecurity differently under different circumstances" [29, p. 2]. Basically, "NIS is work-in-progress" [30, p. 1328], where due to a change in numerous circumstances many countries and organisations acknowledge the need to develop more efficient protection solutions in cyber space and an increase in information security [31].

> "The Directive contributed to improving cybersecurity capabilities at a national level, increased cooperation between Member States, and improved the cyber resilience of public and private entities within the sectors encompassed. However, these improvements seem to be no longer sufficient in light of an expanded threat landscape" [17, p. 1].

"The NIS Directive could be considered a late response to an already exacerbated and well-known problem" [13, p. 11].

Due to the perceived challenges in implementation, the European Commission conducted a comprehensive evaluation study on the scope of the NIS1 Directive. According to the *2020 Commission Staff Working Document Impact Assessment Report Accompanying the document Proposal for a Directive of the European Parliament and of the Council on measures for a high common level of cybersecurity across the Union, repealing Directive (EU) 2016/1148*, in spite of the achievements,

> "the NIS Directive also proved its limitations, falling short of ensuring a fully engaging, coherent and pro-active setting that could guarantee an effective take of shared responsibilities and trust among all relevant authorities and businesses… The NIS Directive revealed inherent weaknesses and gaps that make it incapable of addressing contemporaneous and emerging cybersecurity challenges. These concern, among others, a lack of clarity on the NIS scope, insufficient consideration of the increasing interconnectivity and interdependencies within EU economies and societies, the lack of alignment between security requirements and reporting obligations, a lack of effective incentives for information sharing or operational cooperation among relevant authorities and difference in treatment of comparable businesses across Member States and sectors" [32, p. 11].

The NIS1 Directive did not cover all the sectors that provide key services to the economy and society, was deemed to have granted too wide discretionary powers to Member States to mandate the kinds of cybersecurity and incident reporting requirements for OES, and was not perceived to include effective supervision and enforcement [18, p. 225]. NIS1 "transposition proved to be quite divergent across Member States. This has resulted in an uneven playing field, and insufficient preparedness of those entities in the face of new and evolving cybersecurity challenges" [19, p. 3]. For all of these reasons, it was necessary to adopt a new and updated directive.

The NIS2 Directive was adopted to eliminate weaknesses observed during the evaluation study:

> "the existing capabilities are not sufficient to ensure a high level of security of network and information systems within the Union. Member States have very different levels of preparedness, which has led to fragmented approaches across the Union. This results in an unequal level of protection of consumers and businesses, and undermines the overall level of security of network and information systems within the Union. Lack of common requirements on operators of essential services and digital service providers in turn makes it impossible to set up a global and effective mechanism for cooperation at Union level" [22, recital 5].

NIS1 and NIS2 directives

> "[lay] down measures 'to achieve a high common level of cybersecurity across the Union, with a view to improving the functioning of the internal market'. The main difference is that the NIS1 framework focused on the 'security of network and information systems', whereas the NIS2 Directive focuses on the broader notion of 'cybersecurity' as defined in the Cybersecurity Act. This means that the goal is not just to protect network and information systems, but also 'the users of such systems, and other persons affected by cyber threats'. Given the risks cyberattacks pose to users of ICT systems, this is a welcome scope expansion" [19, p. 5].

The main differences between the NIS1 Directive and NIS2 Directive include:

- **Scope:** Although it primarily excludes small and micro-enterprises, NIS2 encompasses a considerably larger scope than NIS1,

incorporating many new categories of entities – particularly government bodies – and places greater emphasis on digital infrastructure and ICT services. Additionally, NIS2 considerably reduces the discretion of Member States, which should lead to a much more uniform application of its scope across the European Union [19, p. 5; 22, article 1 and 7].

- **Member State obligations:** NIS2 more explicitly elaborates the requirements for Member States' national cybersecurity strategies, thus aiming to achieve a more common level of quality [19, p. 5; 22, article 7].

- **Incident management and response:** NIS2 adds a more efficient obligation for ensuring national large-scale incident management and response [19, p. 5; 22, article 14–17; 16; 17].

- **Reporting obligation:** The reporting obligation has been tightened, given that under NIS1 only very little effective reporting occurred [19, p. 5; 22, article 14 and 16; 16; 17].

- **International coordination:** NIS2 focuses more on enforcing effective coordination between Member States, something that did not happen often under the NIS1 framework [19, p. 5; 22, article 11].

- **Information sharing:** Information sharing is more strongly encouraged [19, p. 5; 22, article 11].

- **Supervision and enforcement:** Supervision and enforcement have been tightened [19, p. 5; 22, article 17].

The NIS2 Directive is a significant piece of legislation that aims to improve the cybersecurity of the European Union. This Act aims to remove wide divergences among Member States (cybersecurity requirements imposed on entities providing services or carrying out activities differed significantly in economic terms among Member States with respect to the type of requirement, their level of detail and the method of supervision; requirements imposed by one Member State differed from, or were even in conflict with, those imposed by another Member State; potentially inadequate design or implementation of cybersecurity requirements in one Member State could have repercussions for the cybersecurity of other Member States, etc.), in particular by setting out minimum rules regarding the functioning of a coordinated regulatory framework, laying down mechanisms for effective cooperation among the responsible

authorities in each Member State, updating the list of sectors and activities subject to cybersecurity obligations and providing effective remedies and enforcement measures which are key to the effective enforcement of those obligations. The Directive contains stricter provisions on the obligations of Member States, essential and important entities, and EU institutions, and emphasises the need for more efficient cooperation. It also sets out the baseline for cybersecurity risk-management measures, and reporting obligations across the sectors that fall within its scope.

4. Discussion

The CER Directive provides a framework for physical and cyber resilience and protection of providers of critical services. The objective of the CER Directive is to remove flaws and strengthen the resilience of critical entities. Critical entities are those that provide basic services which are essential for maintaining important social functions, economic activities, public health and safety, and the environment. The NIS2 Directive extends the scope of implementation to new sectors and stakeholders, strengthens supervision through sanctions and brings about better and more efficient cooperation between Member States. Instead of operators of essential services and digital service providers (from the NIS1 Directive), the NIS2 Directive introduces the categories of essential and important entities. Both directives boost the upgraded foundations of physical and digital security, ensuring a resilient economy and society within each Member State and the European Union as a whole.

Both directives contain many mutual references, describe how Member States should apply them in coordination and cooperation (between the bodies responsible for their implementation) and explain how to avoid an administrative burden beyond that which is necessary to achieve the objectives of both directives. Among others, they envisage interlinkages between cybersecurity and physical security, a coherent approach between these two directives. It is especially important to single out the provision stipulating that entities identified as critical entities under the CER Directive should be considered to be essential entities under the NIS2 Directive [22, article 2, point 3]. Furthermore, it states that each Member State should ensure that its national cybersecurity strategy provides for a policy framework for enhanced coordination between competent authorities within that Member State under both directives in the context of information sharing about risks, cyber threats, and incidents, as well as concerning non-cyber risks, threats and incidents,

and the exercise of supervisory tasks [22, article 7; 23, article 4]. The competent authorities under both directives should cooperate and exchange information in relation to cybersecurity risks, cyber threats and cyber incidents, and non-cyber risks, threats and incidents affecting critical entities, as well as in relation to relevant measures taken by competent authorities [22, article 8; 23, article 9]. All of this strongly implies joint implementation of provisions from both directives and development of common characteristics in strengthening resilience and protection, but also brings common risks that can manifest in multiple normative areas.

The development of cyberspace and information and communication technologies is extremely fast and difficult to regulate, particularly when it needs to be implemented at an EU level and aligned with the vision of EU institutions, the possibilities of Member States, the needs of various markets and economies, and the expectations of manufacturers of different information and communication technologies. These issues give rise to questions related to the transposition and implementation of the above-mentioned documents. That is why it is necessary to continuously study this topic, analyse the current situation and focus on elements that need better or more comprehensive regulation. This is especially important because many actors within and outside the Union (such as the countries that are currently engaged in pre-accession negotiations on full EU membership) have a very different understanding of how best to apply the provisions of the directives, including whether it is even possible to implement a significant part of both directives' provisions.[3]

The NIS2 Directive has some potential weaknesses that could limit its effectiveness. First, the language of the Directive is quite complex and contains a lot of technical detail. This could make it difficult primarily for state institutions, but also for other stakeholders to understand and implement all the necessary requirements in a timely fashion. For example, the Directive entered into force on 16 January 2023, and Member States were given a 21-month deadline for its transposition, until 17 October 2024, by which time they should adopt and publish measures necessary for harmonisation with the Directive. Transposition entails the transfer of rights and obligations from the Directive into national legislation, which involves the adoption of mandatory provisions of national law, or revocation or amendment of existing regulations. The role of the European Commission is decisive in the elaboration of a certain number of measures. Thus, for example, with regard to sector-specific EU legal acts which require essential or important entities to adopt cybersecurity risk-management measures or notify significant incidents,

3 —— The author of this text has been the national contact point for the protection of critical infrastructure and participated in numerous joint EU meetings. Additionally, he was a member of various national working groups for drafting laws and strategies, leading several of them, including the working group for drafting the Law on Critical Infrastructures. Moreover, as an expert, he was engaged by the UN, EU and DCAF to draft laws and by-laws, and implement workshops in the field of critical physical and digital infrastructure protection in the following countries: Bosnia and Herzegovina, Montenegro, Serbia, North Macedonia, Albania and Kosovo. Throughout these experiences, he identified numerous open questions, which he partially addresses in this analysis.

and where those requirements are at least equivalent in effect to the obligations laid down in this Directive, the Commission shall provide guidelines clarifying the application of those measures and requirements by 17 July 2023 [22, article 4]. However, as this research was concluding (at the end of December 2023), this document is still not publicly available (published).

The next example refers to a rather flexible approach to the adoption of a certain number of implementing acts, based on the provision that "the Commission may" adopt them. These are: a) implementing acts laying down procedural arrangements necessary for the functioning of the Cooperation Group [22, article 14]; b) implementing acts laying down the technical and methodological requirements, as well as sectoral requirements, as necessary, of the measures to manage the risks posed to the security of network and information systems which those entities use for their operations or for the provision of their services, and to prevent or minimise the impact of incidents on recipients of their services and on other services [22, article 21]; c) implementing acts further specifying the type of information, format and procedure of a notification, which will ensure that essential and important entities notify about any incident that has a significant impact on the provision of their services [22, article 23]; d) a European cybersecurity certification scheme regarding the use of certain certified ICT products, ICT services and ICT processes [22, article 24].

Additionally, the next example also involves implementing acts to be adopted by the Commission. These are: a) implementing acts laying down the technical and methodological requirements of cybersecurity risk-management measures with regard to DNS service providers, TLD name registries, cloud computing service providers, data centre service providers, content delivery network providers, managed service providers, managed security service providers, providers of online market places, online search engines and social networking services platforms, and trust service providers [22, article 21]; b) implementing acts regarding reporting obligations (specifying the cases in which an incident shall be considered to be significant) of DNS service providers, TLD name registries, cloud computing service providers, data centre service providers, content delivery network providers, managed service providers, managed security service providers, as well as providers of online marketplaces, online search engines and social networking service platforms [22, article 23]. The Commission is required to adopt these acts by 17 October 2024, which is the also deadline for Member States to adopt and publish national measures for harmonisation with the Directive.

The coincidence of these two deadlines should have been avoided, because some countries will wait for the Commission's implementing acts and will not respect the set deadline.

Moreover, it should be noted that the authors of the NIS2 Directive avoided defining the term "crisis" and/or "cyber crisis" and did not lay down the escalation procedure in the event of a cyber crisis. The Member States are responsible for and committed to cooperation within their national framework, and at the EU level via the CSIRTs network, the Cooperation Group, and the European cyber crisis liaison organisation network (EU-CyCLONe), whereby the Commission has its representative in the Cooperation Group, and an observer in the CSIRTs network and the EU-CyCLONe [22, article 13–17]. This arrangement – without a clear explanation of the term "crisis" and/or "cyber crisis" and failing to lay down the escalation procedure in the event of a cyber crisis – represents a serious challenge for implementation of the Directive and efficient management of cyber crises both at the level of Member States and the EU.

Finally, the NIS2 Directive does not specifically consider the growing threat of quantum computing and artificial intelligence. Quantum computers could pose a significant challenge to current cybersecurity measures, and the Directive does not provide any guidance on how to mitigate this threat. Artificial intelligence is mentioned only in the introductory explanations for the adoption of the Directive as a potential means for strengthening the capability and protection of networks and information systems, without specifying any risks that artificial intelligence could create, such as its uncontrolled autonomy.

Unlike the NIS1 Directive and NIS2 Directive, which both focus on the protection of network and information systems within cyberspace, the CER Directive deviated in two ways from Council Directive 2008/114/EC. First, the scope of application is different – whereas Council Directive 2008/114/EC concentrated on the area of security, the CER Directive focuses on the internal market. Second, Council Directive 2008/114/EC addressed critical infrastructure and the CER Directive honed in on critical entities that provide critical services requiring critical infrastructure. Since this document was prepared simultaneously with the NIS2 Directive, both documents share many similarities, which can be positive, but may also lead to some challenges in implementation.

The CER Directive also introduces numerous improvements. While Council Directive 2008/114/EC devoted attention to the procedures for determining critical European infrastructure in the energy and

transport sectors, where disruptions in operation or destruction would have considerable cross-border effects on at least two Member States, and focused exclusively on the protection of such infrastructure, the CER Directive emphasises improved risk assessment, definition and coherence of the roles and duties of critical entities as providers of services which are crucial for the functioning of the internal market of the Union in 11 sectors. Critical entities, with the help of the state, should strengthen their capacity to prevent, protect against, respond to, remain resilient to, mitigate, absorb and recover from incidents that can disrupt the provision of critical services. It should be noted that the number of sectors and subsectors in the CER Directive is significantly higher compared to Council Directive 2008/114/EC (such an approach is also used in the NIS2 Directive compared to the NIS1 Directive), where categories of entities are defined too broadly, which will lead to great challenges and even problems for Member States in the identification and designation of critical entities.

The CER Directive uses the phrase "the Commission may" much less often when laying down the obligations of the Commission for the adoption of implementing acts. It is used only twice, for issues that do not affect transposition and implementation into national legislation (in the first case, the possibility of inviting experts from the European Parliament to attend meetings of the Critical Entities Resilience Group **[23, article 19]**, and in the second, to adopt implementing acts laying down procedural arrangements necessary for the functioning of the Critical Entities Resilience Group **[23]**). However, as in the case of the NIS2 Directive, the CER Directive entered into force on 16 January 2023, and the Member States were given a 21-month deadline for transposition, until 17 October 2024, by which time they must adopt and publish measures necessary for harmonisation with the Directive. The Commission is required to adopt several implementing acts that will enable efficient transposition and implementation into national legislation. The problem is that the final deadline for adoption has only been set for one of them (Risk assessment by Member States), namely, five years after 16 January 2023, whereas several provisions envisage flexibility regarding voluntary adoption and do not set a deadline. These are: a) in cooperation with Member States, to prepare a voluntary common reporting template for reporting on risk assessment of a Member State **[23, article 5]**; b) in cooperation with Member States, to develop recommendations and non-binding guidelines for support to Member States in identifying critical entities **[23, article 6]**; c) upon consultation with the Critical Entities Resilience Group, to adopt non-binding guidelines to facilitate the application of criteria for

determining the significance of negative impact [23, article 7]; d) in cooperation with the Critical Entities Resilience Group, to prepare a joint template for reporting on cross-border cooperation between states; e) upon consultation with the Critical Entities Resilience Group, to adopt non-binding guidelines which further define the technical, security and organisational measures that can be taken as measures for the resilience of critical entities [23, article 13].

Furthermore, as with the NIS2 Directive, the challenges include no mention of "crisis" and lack a detailed elaboration of crisis management escalation procedures. This has been left completely in the hands of Member States, which should develop resilience measures for critical entities to ensure the implementation of risk and crisis management procedures, and protocols and alert routines, but are required to inform the Commission in the event of an incident that has or might have a significant impact on the continuity of the provision of critical services for six or more Member States.

The last challenge refers to the lack of procedural measures related to critical entities built and/or largely managed by EU institutions, whose critical services are used by all Member States. These are critical entities of considerable strategic importance and include: Eurocontrol, a pan-European, civil-military organisation dedicated to supporting European aviation; the Galileo global navigation satellite system; and MeteoAlarm, a European alerting system for extreme weather, etc.

5. Conclusion

With the new package of legislative acts adopted at the end of 2022 and the beginning of 2023, EU institutions attempted to standardise existing practices and challenges in cyberspace, cybersecurity and physical security of network and information systems and critical entities on which business operations in numerous markets depend, as well as the security of states, organisations and individuals. The specific focus of this paper includes two new legislative acts (the NIS2 Directive and the CER Directive), which represent a significant normative improvement and will surely contribute to more efficient measures to strengthen resilience and protection, better cooperation and communication between numerous stakeholders, and less exposure and damage as a result of incidents and irregularities in the functioning of various parts of the system. However, as no perfect regulation exists, the NIS2 Directive and CER Directive have certain weaknesses. This paper addressed the research question posed and demonstrated that these two new

legislative acts have certain flaws that will create challenges for transposition and implementation. Some can be resolved quickly by preparing implementing acts, while the issue of crisis management is more time-consuming and potentially involves a revision of the two documents, or preparation of a supporting document to fill in the existing gaps. In this regard, this research represents a small contribution to the discussion on the protection of the EU's critical infrastructures.

A lot of effort has been invested in the preparation and adoption of both directives, which should be applauded, and they will greatly improve the resilience and protection of network and information systems and critical entities throughout the EU, both individually and collectively with other acts. However, some fear that both directives, particularly the NIS2 Directive, will cause considerable problems in implementation, especially in countries with weak cybersecurity enforcement regimes.

These challenges will likely be exacerbated by the fact that neither directive provides clear guidance on how to implement all of its requirements. Too much flexibility in the adoption of implementing acts by the Commission, which are essential to the transposition and implementation of both directives into national legislations, should have been avoided at all costs. Additional effort was needed to develop said acts and give Member States all the necessary tools for their transposition and implementation. One possibility would have been providing detailed examples and case studies for implementation of all the provisions in the directives.

Both directives list too many sectors, subsectors and categories on the basis of which it is possible to identify and designate essential and important entities (according to the NIS2 Directive), and critical entities (according to the CER Directive). This feels like a too broad approach, where too much room has been left for different interpretations. It is highly probable that too many operators of various facilities, networks and/or systems will be declared essential and important entities, and critical entities, which will lead to challenges in implementation compared to the NIS1 Directive and Council Directive 2008/114/EC. I will provide two examples, one from Europe and the other from the US, which illustrate the problem of identification and designation of critical physical infrastructures (or according to the new conceptualisation – critical entities providing critical services via critical infrastructure). The first example involves the number of identified and designated national critical infrastructures. The available data are very interesting. Here are some of the countries that submitted their

data on critical infrastructures for the purposes of evaluating Council Directive 2008/114/EC: Austria, approx. 400; Czech Republic, approx. 1,900; Estonia, 14; France, 1,438; Hungary, 270; Germany, approx. 1,700; Poland, approx. 550; Portugal, 162; Slovenia, 63 [33]. The research was conducted in 2018 and 2019 and the study was published by the Commission in 2020. Although each country has a certain number of sectors in which it is possible to identity and designate critical infrastructures, the final numbers clearly illustrate a very different understanding of what is critical within each country. The second example involves the number of sectors in which critical infrastructures have been identified and designated in the US, a global leader in regulation of this area. Though the process initially identified a smaller number of sectors, they increased over time to include several thousand facilities, networks and systems designated as critical infrastructures within 16 sectors. Pragmatic Americans realised this was too much and decided to retain all 16 sectors, selecting four that were deemed "more important" than the others and calling them "sectors with lifeline functions": communications, energy, transport and water [34, p. 175]. These two examples show the challenges that arise when it is possible to identify and designate too many elements in too many sectors as critical infrastructure, which consequently leads to problems in implementation, cooperation, coordination and management.

The biggest oversight was the failure to elaborate the issue of crisis and crisis management in both directives. This was left to the Member States and the Union will secure a platform for their cooperation, which is not a satisfactory solution. This is risky for three reasons. First, we all know that the Union is extremely dependent on external energy sources supplied from remote locations, where the majority of transport oil and gas pipelines and shipping routes pass through areas of insecurity and conflict. Second, the initial reaction to the COVID-19 pandemic and the ensuing crisis demonstrated a belated response and the unpreparedness of EU institutions to deal with crisis management. Instead of the European Union managing the crisis on European soil, it was reduced to offering support to Member States. Third, the war in Ukraine has revealed significant discrepancies in points of view and common policies between the EU and Member States, not to mention between the Member States themselves. That is why it is important for the Union to exert stronger leadership in crisis management. The current situation, in which the Union hopes that its Member States will solve crises of a supranational character, with a representative of the Commission as an observer, is not a good solution and a dangerous one because Member States are not capable of this. Instead, the EU should become an active "crisis manager" by addressing all these key issues. There are many obstacles to achieving

this, but with these new legislative acts, an opportunity was lost to adopt a stronger position at the centre of events and resolve potential crisis situations. In addition, the parts of the directives linked to crisis management refer to the exchange of information from operators to competent state institutions and then to the European Commission, with no mention of what the reverse process would look like.

References

[1] The European Commission. (Jul. 24, 2020). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on the EU Security Union Strategy, COM(2020) 605 final. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0605. [Accessed: Nov. 12, 2023].

[2] P. Contreras, "The Transnational Dimension of Cybersecurity: The NIS Directive and Its Jurisdictional Challenges," in *Proceedings of the International Conference on Cybersecurity*, Situational Awareness and Social Media. Springer Proceedings in Complexity, C. Onwubiko, C. et al. Singapore: Springer, 2023, pp. 327–341, doi: 10.1007/978-981-19-6414-5_18.

[3] U. Franke, J. Turell, I. Johannson, "The Cost of Incidents in Essential Services – Data from Swedish NIS Reporting," in *Critical Information Infrastructures Security. CRITIS 2021. Lecture Notes in Computer Science*, vol. 13139, D. Percia David, A. Mermoud, T. Maillart, Eds. Cham: Springer, 2021, pp. 116–129, doi: 10.1007/978-3-030-93200-8_7.

[4] A. Mishra, Y. I. Alzoubi, M. J. Anwar, A. Q. Gill, "Attributes impacting cybersecurity policy development: An evidence from seven nations," *Computers & Security*, vol. 120, 2022, pp. 1–23, doi: 10.1016/j.cose.2022.102820.

[5] S. Maesschalck, V. Giotsas, B. Green, N. Race, "Don't get stung, cover your ICS in honey: How do honeypots fit within industrial control system security," *Computers & Security*, vol. 114, pp. 1–25, 2022, doi: 10.1016/j.cose.2021.102598.

[6] M. Mirtsch, K. Blind, C. Koch, G. Dudek, "Information security management in ICT and non-ICT sector companies: A preventive innovation perspective," *Computers & Security*, vol. 109, pp. 1–23, 2021, doi: 10.1016/j.cose.2021.102383.

[7] D. Polverini, F. Ardente, I. Sanchez, F. Mathieux, P. Tecchio, L. Beslay, "Resource efficiency, privacy and security by design: A first experience on enterprise servers and data storage products triggered by a policy process," *Computers & Security*, vol. 76, pp. 295–310, 2018, doi: 10.1016/j.cose.2017.12.001.

[8] C. Banasiński, M. Rojszczak, "Cybersecurity of consumer products against the background of the EU model of cyberspace protection," *Journal of Cybersecurity*, vol. 7, no. 1, pp. 1–15, 2021, doi: 10.1093/cybsec/tyab011.

[9] H. Kavak, J. J. Padilla, D. Vernon-Bido, S. Y. Diallo, R. Gore, S. Shetty, "Simulation for cybersecurity: state of the art and future directions," *Journal of Cybersecurity*, vol. 7, no. 1, pp. 1–13, 2021, doi: 10.1093/cybsec/tyab005.

[10] S. Varga, J. Brynielsson, U. Franke, "Cyber-threat perception and risk management in the Swedish financial sector," *Computers & Security*, vol. 105, pp. 1–18, 2021, doi: 10.1016/j.cose.2021.102239.

[11] J. D. Michels, I. Walden, "How Safe is Safe Enough? Improving Cybersecurity in Europe's Critical Infrastructure Under the NIS Directive," Queen Mary School of Law Legal Studies, Research Paper No. 291/2018, pp. 1–47. [Online]. Available: https://ssrn.com/abstract=3297470. [Accessed: Nov. 18, 2023].

[12] T. Aleksandrowicz, "The Act on the National Cybersecurity System as an Implementation of the NIS Directive," *Internal Security*, vol. 12 no. 1, pp. 179–193, 2020, doi: 10.5604/01.3001.0014.3196.

[13] D. Markopoulou, V. Papakonstantinou, P. de Hert, "The new EU cybersecurity framework: The NIS Directive, ENISA's role and the General Data Protection Regulation," *Computer Law & Security Review,* vol. 35, no. 6, pp. 1–11, 2019, doi: 10.1016/j.clsr.2019.06.007.

[14] M. D. Cole, S. Schmitz-Berndt, "The Interplay between the NIS Directive and the GDPR in a Cybersecurity threat landscape," University of Luxembourg Law Working Paper No. 2019–017, pp. 1–20. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3512093. [Accessed: Dec. 3, 2023].

[15] S. Schmitz-Berndt, S. Schiffner, "Don't tell them now (or at all) – responsible disclosure of security incidents under NIS Directive and GDPR," *International Review of Law, Computers & Technology*, vol. 35, no. 2, pp. 101–115, 2021, doi: 10.1080/13600869.2021.1885103.

[16] S. Schmitz-Berndt, „Refining the Mandatory Cybersecurity Incident Reporting Under the NIS Directive 2.0: Event Types and Reporting Processes," in *Proceedings of the International Conference on Cybersecurity, Situational Awareness and Social Media*. Springer Proceedings in Complexity. C. Onwubiko et al. Singapore: Springer, 2023, pp. 343–351, doi: 10.1007/978-981-19-6414-5_19.

[17] S. Schmitz-Berndt, "Defining the reporting threshold for a cybersecurity incident under the NIS Directive and the NIS2 Directive," *Journal of Cybersecurity*, vol. 9, no. 1, pp. 1–11, 2023, doi: 10.1093/cybsec/tyad009.

[18] T. Sievers, "Proposal for a NIS directive 2.0: companies covered by the extended scope of application and their obligations," *International Cybersecurity Law Review*, vol. 2, pp. 223–231, 2021, doi: 10.1365/s43439-021-00033-8.

[19] N. Vandezande, „Cybersecurity in the EU: How the NIS2-directive stacks up against its predecessor," SSRN, pp. 1–16, 2023. [Online]. Available: https://ssrn.com/abstract=4383118. [Accessed: Dec. 10, 2023].

[20] A-V. Dragomir, "What's new in the NIS2 Directive Proposal Compared to the Old NIS Directive," *SEA – Practical Application of Science*, vol. 9, no. 27, pp. 155–162, 2021 [Online]. Available: https://seaopenresearch.eu/Journals/articles/SPAS_27_1.pdf. [Accessed: Nov. 30, 2023].

[21] S. Schmitz-Berndt, P. G. Chiara, "One step ahead: mapping the Italian and German cybersecurity laws against the proposal for a NIS2 directive," *International Cybersecurity Law Review*, no. 3, pp. 289–311, 2022, doi: 10.1365/s43439-022-00058-7.

[22] The European Parliament and the Council of the European Union. (Dec. 14, 2022). Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS2 Directive) [Online]. Available: https://eur-lex.europa.eu/eli/dir/2022/2555/oj. [Accessed: Aug. 12, 2023].

[23] The European Parliament and the Council of the European Union. (Dec. 14, 2022). Directive (EU) 2022/2557 of the European Parliament and of the Council of 14 December 2022 on the resilience of critical entities and repealing Council Directive 2008/114/EC. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022L2557&qid=1692286376725. [Accessed: Aug. 12, 2023].

[24] The European Parliament and the Council of the European Union. (Dec. 14, 2022). Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014, (EU) No 909/2014 and (EU) 2016/1011. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2554. [Accessed: Aug. 13, 2023].

[25] The Council of the European Union. (Dec. 8, 2008). Council Directive 2008/114/EC of 8 December 2008 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32008L0114. [Accessed: Aug. 14, 2023].

[26] The European Commission, Directorate-General for Migration and Home Affairs. *Evaluation study of Council Directive 2008/114 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection, Executive summary*, 2019, doi: 10.2837/353895.

[27] The European Commission, Directorate-General for Migration and Home Affairs. *Evaluation study of Council Directive 2008/114 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection – Final report*, 2019, doi: 10.2837/864404.

[28] The European Parliament and the Council of the European Union. (Jul. 6, 2016). Directive (EU) 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. [Online]. Available: https://eur-lex.europa.eu/eli/dir/2016/1148/oj. [Accessed: Aug. 14, 2023].

[29] V. Papakonstantinou, "Cybersecurity as praxis and as a state: The EU law path towards acknowledgement of a new right to cybersecurity?" *Computer Law & Security Review*, vol. 44, pp. 1–15, 2022, doi: 10.1016/j.clsr.2022.105653.

[30] O. Michalec, S. Milyaeva, A. Rashid, "Reconfiguring governance: How cyber security regulations are reconfiguring water governance," *Regulation & Governance*, vol. 16, no. 4, pp. 1325–1342, 2022, doi: 10.1111/rego.12423.

[31] E. K. Szczepaniuk, H. Szczepaniuk, T. Rokicki, B. Klepacki, "Information security assessment in public administration," *Computers & Security*, vol. 90, pp. 1–11, 2020, doi: 10.1016/j.cose.2019.101709.

[32] The European Commission. *Commission Staff Working Document Impact Asessment Report Accompanying the document Proposal for a Directive of the European Parliament and of the Council on measures for a high common level of cybersecurity across the Union, repealing Directive (EU) 2016/1148, SWD/2020/345 final – part 1/3,* 2020. [Online]. Available: https://eur-lex.europa.eu/resource.html?uri=cellar:d51e4bbb-3fa8-11eb-b27b-01aa75ed71a1.0001.02/DOC_1&format=PDF. [Accessed: Dec. 13, 2023].

[33] The European Commission, Directorate-General for Migration and Home Affairs. *Evaluation study of Council Directive 2008/114 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection, Annex II*, 2020. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/71835078-b043-11ea-bb7a-01aa75ed71a1/language-en/format-PDF/source-search. [Accessed: Aug. 21, 2023].

[34] R. Mikac, I. Cesarec, R. Larkin, *Critical infrastructure: A platform for the successful development of the security of nations*. Zagreb: Jesenski i Turk (in Croatian), 2018.

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Regulating Deep Fakes in the Artificial Intelligence Act

Mateusz Łabuz | Ministry of Foreign Affairs of the Republic of Poland, Chemnitz University of Technology, Germany, ORCID: 0000-0002-6065-2188

Abstract

The Artificial Intelligence Act (AI Act) may be a milestone in the regulation of artificial intelligence by the European Union. The regulatory framework proposed by the European Commission has the potential to serve as a global benchmark and strengthen the position of the EU as one of the main players on the technology market. One of the components of the draft regulation are the provisions on deep fakes, which include a relevant definition, risk category classification and transparency obligations. Deep fakes rightly arouse controversy and are a complex phenomenon. When leveraged for negative purposes, they significantly increase the risk of political manipulation, and at the same time contribute to disinformation, undermining trust in information and the media. The AI Act may strengthen the protection of citizens against some of the negative consequences of misusing deep fakes, although the impact of the regulatory framework in its current form will be limited due to the specificity of their creation and dissemination. The effectiveness of the provisions will depend not only on enforcement capabilities, but also on the precision of phrasing provisions to prevent misinterpretation and deliberate abuse of exceptions. At the same time, the AI Act will not cover a significant portion of deep fakes, which, due to the malicious intentions of their creators, will not be subject to the transparency obligations. This study analyses provisions related to deep fakes in the AI Act and proposes improvements that will take into account the specificity of this phenomenon to a greater extent.

Keywords

deep fakes, Artificial Intelligence Act, AI Act, regulations, European Union, transparency obligations, disclosure rules

Received: 14.09.2023

Accepted: 19.10.2023

Published: 27.10.2023

Cite this article as:
M. Łabuz "Regulating Deep Fakes in the Artificial Intelligence Act," ACIG, vol. 2, no. 1, 2023, DOI: 10.60097/ACIG/162856

Corresponding author:
Mateusz Łabuz, Ministry of Foreign Affairs of the Republic of Poland, Chemnitz University of Technology, Germany
ORCID: 0000-0002-6065-2188; E-MAIL: mateusz.labuz@msz.gov.pl

1. Introduction

The proposal for Regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (AI Act)[1], introduced by the European Commission in April 2021, is intended to be one of the key elements in positioning the European Union to regulate the dynamic development of artificial intelligence (AI). Creating a legal framework for AI will not only allow the EU to face numerous legal, political, economic and social challenges, but will also put it in a privileged position in the global competition to set regulatory standards [1, 2], and perhaps even to 'serve as a benchmark for other countries' [3]. The AI Act addresses the risks associated with certain uses of technology and aims to achieve 'the development of an ecosystem of trust by proposing a legal framework for trustworthy AI' [4]. Creating 'an ecosystem of trust' is an ambitious task, which requires internal consistency and a great sense in using legal terms, so they are not contested or interpreted in a way that is incongruent with the spirit of the regulation.

1 — As of the date this paper was written (August 2023), the AI Act was still being negotiated.

Deep fakes, which first began to appear in 2017, are a relatively well-described phenomenon and have been the subject of numerous analyses that, among others, extensively described the various ways they are used to inflict harm [5–8]. The potential risks of the misuse of deep fakes include the spread of fake news and disinformation, election manipulation, creation of non-consensual pornographic content, defamation, discredit and ridicule of individuals, including political opponents, undermining trust in traditional media messages, distortion of reality, impairment of political engagement within society, undermining of the epistemic quality of debate and thus democratic discourse, threats to the stability of economic systems, spread of hate speech and strengthening gender inequalities, as well as psychological harm to individuals or vulnerable groups [9–11].

This non-exhaustive enumeration does not fully reflect the specificity of the phenomenon. One should not forget that deep fakes find many positive applications in the media, education, leisure and healthcare [12]. Therefore, they should not be demonised wholesale and every legal solution should take into account the diversity of uses and consequences related to the creation and dissemination of deep fakes in their various forms [13].

To date, deep fakes have somehow eluded basic legislation and rules governing their use have mostly been taken from provisions of civil, tort, criminal or copyright law [14]. The first attempts to regulate deep fakes in more specific legal acts should be observed with

interest, especially in terms of their implementation and impact on social and political processes. The EU can play a constructive role in this process, not only by referring to deep fakes in the AI Act, but also by using that leverage to introduce stricter countermeasures. The development of technology and frequently reported misuses require deep fakes to be directly regulated and, if necessary, forbidden if they directly violate the rights of third parties [16]. However, one should be realistic – even outright bans would not be completely effective since most deep fakes are meant to deceive recipients and circumvent legal, technical and social safeguards [17]. It is also necessary to consider how the law can protect the basic values of democracy from malicious and non-malicious uses of deep fakes, while preserving fundamental rights, including freedom of speech [18].

The AI Act refers to deep fakes explicitly, introducing a definition of the term, basic transparency and disclosure rules, and assigns deep fakes to the 'specific' or 'limited risk' quasi-category of AI systems [19, 20]. Some of the proposals introduced in 2021 by the Commission were rightly criticised by experts, who pointed to an insufficient legal regime, or underestimation of the seriousness of threats stemming from the creation and dissemination of deep fakes [21–23].

The amendments proposed by the European Parliament in June 2023 [24] have the potential to eliminate some of the deficits in the Commission's original proposal and can be generally deemed as a step in the right direction. At the same time, there are still shortcomings that should be addressed as part of further negotiations in order to create a coherent, although quite general in nature, legal framework for regulating deep fakes. However, it seems crucial to verify whether the proposed solutions will create an effective framework for combating deep fakes, which, in light of previous cases, seems doubtful and may force the EU to quickly examine and adjust its approach to their regulation.

The EU must ensure internal consistency, so that the definitions and solutions proposed in various documents are complementary and do not lead to misinterpretation or discrepancies. At the same time, deep fakes, due to their complexity and the cascading effects of their misuse [25], are a phenomenon that must be taken into account in more specific acts, which paves the way to further discussion on enforcement, liability and penalisation [26].

This study primarily serves to highlight the issue of deep fakes in light of the AI Act and is part of the current debate [9, 27–30] on the risks associated with the dissemination of technology that enables the

creation of hyper-realistic but fake synthetic media [10], which are increasingly difficult to distinguish from real ones [31, 32].

The aim of the study is to assess the current state of the AI Act in regard to deep fakes, as well as to draw attention to the shortcomings of the proposal. As already mentioned, deep fakes cannot be categorised unequivocally due to their multitude of applications. At the same time, it should be emphasised that they play an increasingly important role in entrenching digital disinformation [33, 34] and negatively affect many spheres of life [6, 9]. In some cases, they might directly threaten democracy, free elections and the information ecosystem, undermine trust in the media, or lead to the victimisation of individuals, especially women [9, 35]. The comments made by the author may also serve as a signpost for policymakers who, regardless of EU regulations, sooner or later will have to face the problem of deep fakes at the level of national legislation.

As the AI Act is still being negotiated, further changes to the substance of the regulation are possible, which might make it possible to eliminate deficits or shortcomings in already implemented measures.[2] Due to possible changes in the regulation, this study might become obsolete when the AI Act is adopted, which is a significant limitation. Nevertheless, analysis of the draft proposals and criticism of selected solutions can provide additional input for discussions on creating a regulatory framework in relation to deep fakes, which increases the paper's topicality and applicability.

2 —— The analysis was based on the proposals from April 2021 (the European Commission's proposal) and June 2023 (European Parliament amendments).

2. Definition scope

A holistic approach to the issue of deep fakes requires, first of all, the introduction of a legal definition of this term. The AI Act may be a point of reference for further work and legislation in this regard, which makes the EU's approach to the definition of key importance.

The Commission's proposal [4] referred only to a quasi-definition of 'deep fake'. Although the most relevant Article 3 contained definitions of terms used by the AI Act, deep fakes were not included in the list. The description of a deep fake was inserted into Article 52(3), which was supposed to set out transparency obligations for certain AI systems:

> *Users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would*

> *falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated.*

Article 52(3) was later amended by the European Parliament to expand the range of the quasi-definition and introduce stricter transparency obligations:

> *Users of an AI system that generates or manipulates text, audio or visual content that would falsely appear to be authentic or truthful and which features depictions of people appearing to say or do things they did not say or do, without their consent ('deep fake'), shall disclose in an appropriate, timely, clear and visible manner that the content has been artificially generated or manipulated, as well as, whenever possible, the name of the natural or legal person that generated or manipulated it.*

The scope of this provision will be discussed at a later stage because the key change in the European Parliament's amendments in regard to defining deep fakes is the addition of a new point 44d in Article 3(1), which introduces the legal definition of the term and should be treated as a point of reference:

> *'deep fake' means manipulated or synthetic audio, image or video content that would falsely appear to be authentic or truthful, and which features depictions of persons appearing to say or do things they did not say or do, produced using AI techniques, including machine learning and deep learning.*

The addition of point 44d in Article 3(1) allows us to extract the four most important aspects of the definition of deep fakes:

1. technical, relating to the method of creation (manipulated or synthetic content, produced using AI techniques, including machine learning and deep learning);
2. typological, relating to the form of media that was used (audio, image or video content);
3. subjective, referring to the subject/object of depiction (features persons);
4. effectual, relating to the manner and effect of depiction (falsely appears to be authentic or truthful; appearing to say or do things they did not say or do).

Only meeting all aspect criteria together constitutes a deep fake. The survey conducted by A. Fernandez [21] to establish the elements of commonly used definitions of deep fakes resulted in the recognition of two mandatory features agreed on by scholars: 1) intervention by AI (which overlaps with the technical aspect); 2) the potential to deceive (which overlaps with the effectual aspect). Referring to A. de Ruiter [27], Fernandez considered the deceptive effect as a 'by-product of the creator's intent'. This approach seems only partially correct, as the very nature of a deep fake is based on the presentation of a false or distorted reflection of reality and thus implies intent to deceive recipients.

In general the decision to introduce a legal definition of 'deep fake' within the AI Act should be assessed positively. The Parliament's deletion of the wording 'to a person' is an unequivocally positive development in comparison to the Commission's phrasing of Article 52(3), because it expands the range of entities that can be targeted by audio or visual forgery, specifies the technical aspects, and makes explicit mention of machine learning and deep learning technologies. One could rightly argue that confirming whether content constituted a deep fake would require proving that an AI system was used to generate it. The degree of technological advancement with respect to tools used to create deep fakes is so high that unambiguous evidence that AI was used to generate content may be difficult or even impossible to obtain [21]. However, the definition does not seem to offer a reasonable alternative for the technical aspect and only the practical functioning of the provisions will reveal whether the classification of materials as deep fakes is rendered impossible by an inability to prove the use of AI.

To this extent, the definition in point 44d of Article 3(1) extends the scope of the quasi-definition included in the original Article 52(3) proposed by the Commission.

Intervention in the subjective aspect, which narrows the scope of subjects/objects depicted [24] with the wording *which features depictions of persons appearing to say or do things they did not say or do* is a negative development. Reference to objects, places or other entities or events that appeared in the original Commission proposal have been erroneously deleted, which limits the possibility of classifying content as a deep fake.

Meanwhile, there are deep fakes that do not depict people, but have proven to be effective tools in significantly influencing reality. In May 2023, an image deep fake depicting an explosion near the

Pentagon was disseminated via social media, leading to short-term losses on the New York Stock Exchange. According to Bloomberg, it was 'possibly the first instance of an AI-generated image moving the market' [36]. It is also possible to imagine images of natural disasters, military equipment, war damage, or desecration of religious symbols that do not include people [37]. Each one could serve as an inflammatory spark, leading to social unrest, or mobilisation of specific groups, and contribute to disinformation [38]. Some researchers are already warning of 'deep fake geography', which refers to falsification of cartographical data, including satellite images [39, 40]; some states have allegedly already used such images for the purpose of sophisticated disinformation [41].

Although some scenarios are for now only the subject of speculation, they are already being discussed by researchers, who are trying to raise awareness among policymakers. For this reason, extending the subjective scope of the definition of 'deep fake' and using the earlier proposal of the Commission seems advisable. Interestingly, the European Parliament's Committee on Industry, Research and Energy, in its opinion of June 2022 [42], proposed a deep fake definition that referred to *material that gives an authentic impression, in which events appear to be taking place, which never happened,* completely omitting any remarks concerning 'persons'.

In Recital 70 of the initial draft of the AI Act [4], which has not been amended by the European Parliament, the Commission referred

> *to certain AI systems intended to interact with natural persons or to generate content [that] may pose specific risks of impersonation or deception irrespective of whether they qualify as high-risk or not. And later: users, who use an AI system to generate or manipulate image, audio or video content that appreciably resembles existing persons, places or events and would falsely appear to a person to be authentic, should disclose that the content has been artificially created or manipulated by labelling the artificial intelligence output accordingly and disclosing its artificial origin.*

Though deep fakes were not explicitly mentioned in Recital 70, they definitely match the description, which again features an extended subjective scope.

Unfortunately, there are significant differences between Recital 70, Article 3(1) point 44d and Article 52(3). These discrepancies need to be clarified in future.

Article 52(3) [24] leads to even more confusion as to whether the typological aspect was extended by the European Parliament to include text forms. Although deep fakes have no agreed-upon technical or typological definition [43], some concepts are circulating among scholars. While in the majority of the analysed studies, including reports from the European Union Agency for Cybersecurity (ENISA), EUROPOL, NATO, and an AI glossary by Brookings, the definitions are narrowed down and explicitly mention audio, image or video content [34, 44–47], some researchers also mention deep fakes in text form [28, 48–50], or are even concentrating on developing deep fake text detection methods [51].

At this stage, the AI Act does not completely resolve the problem of qualifying textual deep fakes. The European Parliament decided not to include text form of deep fakes within the scope of the definition included in Article 3(1). On the other hand, the amended Article 52(3), referring to transparency obligations, creates ambiguity, as the Commission's proposal was supplemented with the term 'text'. In the further part of the provision it was indicated that this might also refer to deep fakes. The literal understanding of the provision suggests that the scope of definition contained in Article 3(1) has been extended with respect to the typological aspect. Undoubtedly, appropriate disclosure rules should also apply to AI-generated or AI-manipulated texts, but they do not necessarily have to be qualified as deep fakes. This ambiguity needs to be clarified in future.

It is worth mentioning that the issue of extending the definition to include text deep fakes was raised by the European Parliament's Committee on Culture and Education [52], whose proposal for amendments of June 2022 referred to a deep fake as: *manipulated or synthetic audio, visual or audiovisual content, text or scripts which feature persons purported to be authentic and truthful.* A similar position (advocating for inclusion of deep fakes in text form) was taken by Mesarčík et al. [22] in a critical analysis of the AI Act proposal, though this study did not contain a proper rationale for such an inclusion.

In the author's opinion, the EU needs to either clearly include text deep fakes in the definition in Article 3(1), or clearly distinguish the text form from deep fakes, focusing only on audio and visual content in Article 52(3), and possibly create an additional provision for AI-generated and AI-manipulated texts. The latter seems to be the solution that would better match the analyses carried out by the majority of researchers.

Another problem in defining deep fakes in EU legal acts is the consistency of the proposed solutions. Deep fakes rarely receive an explicit

legal definition; more often regulations can be derived from specific formulations relating to phenomena that are similar or identical to deep fakes. The AI Act can serve as a benchmark for other legal acts, which in turn requires the consistent use of one term and one qualification. If the AI Act introduces the legal definition of deep fakes, other definitions or quasi-definitions/descriptions must cover the same scope.

One could identify an example of another definition proposed by the EU within the Proposal for a Directive of the European Parliament and of the Council on combating violence against women and domestic violence that refers to the production and dissemination of non-consensual deep porn[3] [54]. According to some estimations, deep fakes of a pornographic nature might constitute more than 90% of all deep fakes circulating on the internet, which clearly shows the scale of the problem [53] and explains the rationale behind including deep fakes into this particular directive in Recital 19:

> *The offence should also include the non-consensual production or manipulation, for instance by image editing, of material that makes it appear as though another person is engaged in sexual activities, insofar as the material is subsequently made accessible to a multitude of end-users, through information and communication technologies, without the consent of that person. Such production or manipulation should include the fabrication of 'deepfakes', where the material appreciably resembles an existing person, objects, places or other entities or events, depicting sexual activities of another person, and would falsely appear to others to be authentic or truthful.*

Regardless of the fact that the creation of deep porn materials should become a criminal offence and the limitation be of a contextual nature, inconsistency in the use of terms draws attention. Spelling discrepancies ('deepfakes' and 'deep fakes'[4]) are not as significant as the varying scope of definitions. It might be surprising that, only in this case, when the depiction of existing persons seems to be of importance due to the nature of deep porn [35], other elements being a part of the subjective aspect are also explicitly mentioned.

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (DSA) [56] does not offer any legal definition of deep fakes but it apparently refers to that phenomenon in Article 35(1). While discussing 'mitigation of risks', DSA points to

3 —— Altered material of a sexual or pornographic nature, depicting people whose faces were superimposed on visual or audiovisual content [53].

4 —— With regard to spelling, one can also note the notation used by experts from the Panel for the Future of Science and Technology (STOA), who consistently used the term 'deep-fakes' in the report on the draft AI Act [55].

the obligations of providers of very large online platforms and very large online search engines, who should put in place reasonable, proportionate and effective mitigation measures. Such measures may include, where applicable:

> *k) ensuring that an item of information, whether it constitutes a generated or manipulated image, audio or video that appreciably resembles existing persons, objects, places or other entities or events and falsely appears to a person to be authentic or truthful, is distinguishable through prominent markings when presented on their online interfaces, and, in addition, providing an easy to use functionality which enables recipients of the service to indicate such information.*

In that case, attention should be drawn primarily to the extended subjective aspect, as the definition covers persons, objects, places or other entities or events. The issue of complementarity and potential strengthening of DSA provisions by the AI Act will be discussed later in the study.

A potential legal act that might in future also include a reference to deep fakes, due to their possible malicious applications in shaping political reality, influencing elections and causing risk of reputational harm to individuals [6], is the Proposal for a Regulation of the European Parliament and of the Council on the transparency and targeting of political advertising [57]. At the moment, this regulation does not directly refer to deep fakes. The table below summarises different definitions or references to deep fakes in EU legal acts.

Table 1. Different definitions and descriptions of deep fakes in the EU legal acts.

The AI Act proposal – European Commission, Article 52(3) (April 2021)	*an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake')*
The AI Act Proposal – European Parliament, Article 3(1) point 44d (June 2023)	*'deep fake' means manipulated or synthetic audio, image or video content that would falsely appear to be authentic or truthful, and which features depictions of persons appearing to say or do things they did not say or do, produced using AI techniques, including machine learning and deep learning*

The AI Act Proposal – European Parliament, Article 52(3) (June 2023)	*an AI system that generates or manipulates text, audio or visual content that would falsely appear to be authentic or truthful and which features depictions of people appearing to say or do things they did not say or do, without their consent ('deep fake')*
Directive on combating violence against women and domestic violence, Recital 19 (March 2022)	*'deepfakes', where the material appreciably resembles an existing person, objects, places or other entities or events, depicting sexual activities of another person, and would falsely appear to others to be authentic or truthful*
Regulation on a Single Market For Digital Services (DSA), Article 35(1k) (October 2022)	*an item of information, whether it constitutes a generated or manipulated image, audio or video that appreciably resembles existing persons, objects, places or other entities or events and falsely appears to a person to be authentic or truthful*

3. Qualification – specific risk category?

Any classification of deep fakes within a risk category should first take into account the possible uses of the technology that definitely may vary [12]. Deep fakes should not only be considered as a dangerous form of audio and visual manipulation, as there are many positive applications of the technology. They should not be described as something 'inherently morally wrong' and the technology itself should rather be considered 'neutral' [13, 27]. It is the use of deep fakes that gives them a certain dimension, and the objectives behind their creation or dissemination that put them into a specific context. The aforementioned elements of the definition do not comprise the aspect of contextuality, which often determines their harmfulness, and thus can be a key factor in risk assessment.

Strong emphasis on the negative uses of deep fakes has given them a bad reputation. There is a possibility that their excessive demonisation will lead to inappropriate risk assessment, or will undermine significant scientific and technological progress achieved with the use of deep fakes. Excessive interventionism, even if motivated by the protection of higher goods, can significantly limit technological development, and thus the competitiveness of the EU. Therefore any regulatory framework must be well-balanced. Discussing the positive uses of deep fakes is beyond the scope of this study, but it is worth noting that the term itself 'now carries negative connotations, potentially causing hesitancy or scepticism when discussing its legitimate research applications' [58].

Some authors believe that deep fakes are not a neutral technology, and that their history basically began with the creation of pornographic content, which clearly shows its original, highly disturbing objectives [59]. EUROPOL experts [60] estimate that most deep fakes are disseminated with malicious intent. Additionally, an intrinsic feature of deep fakes is that they increase confusion by blurring the boundaries between the authentic and the inauthentic and make it difficult to distinguish what is fact and what is fiction [59], significantly enhancing the potential for digital disinformation [25, 33]. Therefore, while assessing the potential applications of deep fakes, one should take into account the general negative consequences they cause within the information space, including undermining trust in information or the media [33, 61, 62].

The AI Act introduces a gradation of three basic risk categories: (i) unacceptable risk, (ii) high risk, and (iii) low or minimal risk. A detailed discussion on the legitimacy of such a division goes beyond the scope of this study, but the general idea of risk regulations is to prevent risk by reducing the probability of its occurrence [63]. It should be noted that in some respects the categorisation proposed in the AI Act is 'illusory and arbitrary' and does not apply to the entire 'AI lifecycle', which excludes or does not entirely cover the harmful forms of use of some systems [64]. The very general references (amended Recital 4 of the AI Act [24]) to the societal harm that some systems pose do not help to achieve clarity and certainty of the categorisation [65]. The AI Act does not provide a clear rationale for classifying deep fakes into any of the categories.

Pursuant to the AI Act, deep fakes were not qualified within the first two categories, so they should be automatically considered a low or minimal risk AI system. However, Title IV of the AI Act takes into account the specific risks of manipulation that some AI systems pose and thus introduces additional transparency obligations for specific AI systems. Deep fakes were enumerated among them and covered within the scope of the aforementioned Article 52(3) of the AI Act.

Deep fakes (and chatbots, pursuant to Article 52) must be treated as exceptions within the three-risk-categories system introduced by the AI Act [66] and might be classified as a 'specific risk' or 'limited risk' AI system [19, 20]. Therefore they form a separate quasi-category [67, 68].

Initially, in Recital 38 of the AI Act [4], which enumerates some high-risk systems, the Commission pointed out: *in view of the nature of the activities in question and the risks relating thereto, those high-risk AI*

systems should include in particular AI systems intended to be used by law enforcement authorities (...) to detect 'deep fakes'. According to the qualification made by the Commission, AI systems intended for use by law enforcement authorities to detect deep fakes were included in the list of High-Risk AI Systems (Annex III).

It should then be noted that originally deep fakes were classified into the third category or quasi-category (low or minimal risk, or specific or limited risk due to transparency obligations), while in the Commission's proposal deep fake detection systems were placed in the second category (high-risk). From the beginning, this qualification discrepancy gave rise to astonishment [21, 22, 69]. The misclassification was unconvincingly justified by the assumption that the former are used by the private sector, while the latter would be mainly in the hands of the public sector. Researchers from the European Parliamentary Research Service concluded: 'It is surveillance by the state individuals need protection from' [55]. This justification definitely loses to a practical approach to threats, primarily due to the malicious uses of deep fakes and the necessity to introduce efficient state countermeasures.

The European Parliament [24] effected a key change in this respect by making two deletions regarding deep fake detection systems – in Recital 38 and from the list of High-Risk AI Systems (Annex III) added to the AI Act. This is a direct and rational response to expert reservations and a common-sense approach – assigning a higher risk category to a technology that is supposed to protect against abuses with the use of technology classified in a lower risk category does not make sense. However, reservations may be made to paragraph 1 point 6d of Annex III, where the following are listed among high-risk AI systems: *AI systems intended to be used by or on behalf of law enforcement authorities, or by Union agencies, offices or bodies in support of law enforcement authorities to evaluate the reliability of evidence in the course of investigation or prosecution of criminal offences.* Deep fake detection systems could be indirectly included in the list [70], whereas their use in verifying the veracity of evidence has the potential for growth and is even recommended to ensure evidence integrity [60]. Experts consistently emphasise the importance of deep fake detection tools for counteracting deep fakes of a malicious nature, strengthening the capacity of law enforcement authorities, or protecting judicial proceedings [12, 60, 71]. These countermeasures will probably play an increasingly important role in the face of a growing number of crimes involving the malicious use of deep fakes (extortion, impersonation, financial fraud, forging evidence).[5] Separate issues are the effectiveness and credibility of the detection tools, as well as ensuring fair access to their

5 —— Deep fakes or their appearance in the information space have already been used during court proceedings to provide evidence (case in the UK during a custody battle), or to create a specific line of defense (the so-called 'deep fake defense') by claiming that evidence was fabricated [72], [73]. They also strengthen the 'liar's dividend', allowing depicted persons to claim that real content is in fact fake [6].

use, which should be an element of risk assessment that takes into account trustworthiness.

It is still necessary to return to the basic qualification of deep fakes in risk categories. Some researchers questioned the classification of deep fakes within the low or minimal risk category from the beginning, postulating their inclusion in the high-risk category [22] or reconsidering initial qualification [74]. The proposals were based on the correct assumption that counteracting deep fakes only through transparency obligations misses an important aspect of audiovisual content's manipulative capabilities. Mesarčík et al. [22] advocated for consistency in qualification rules for high-risk systems and accused the Commission of lacking a rationale in the case of deep fakes, the harmfulness of which might directly violate selected fundamental rights. In addition, they indicated the lack of a definition of inappropriate uses of deep fakes. All these objections seem valid, but difficult to grasp due to the high contextualisation of deep fakes and diversification of their applications.

A group of scientists conducting research on the harmful uses of deep fakes [13] indicated that 'manipulations may exhibit different risk levels and the risk level highly depends on the type of specific applications and somewhat subjectively depending on the actual use case'. This is an extremely important observation that relates in the first place to the various purposes behind the creation and dissemination of deep fakes and the contextuality of deep fakes as information carriers [75]. R. T. Toparlak [16] rightly noted that 'the wide range of applications means some deep fakes are going to be high-risk, while others are completely harmless'.

Theoretically, it is the objectives and the appropriate context of a particular deep fake that should determine its qualification into a risk category. In fact, they could be divided into many subgroups, depending on their form and purpose of use. This would of course give rise to problems of interpretation [76], which would have to be resolved on a case-by-case basis. However, the division would better reflect the specificity of the use of deep fakes and their destructive impact on the information ecosystem, individuals and society.

Assessing the harmfulness of deep fakes or the purposes of their creation and dissemination on a case-by-case basis seems rather unrealistic, which in turn undermines the legitimacy of the exception-based or multi-qualification risk system. Some authors [59] rightly noted that the scale of production of audio and visual materials is so large that it exceeds the verification capabilities of any institution, and

the verification itself would most likely have to be based on human review [21].

The answer to the question of whether deep fakes should qualify as a high-risk AI system is not clear. In light of descriptions of the high-risk AI systems category presented in the AI Act, one can have reasonable doubts whether deep fakes fail to meet at least some of the criteria. In the Explanatory Memorandum of the AI Act, it was noted that high-risk AI systems pose significant risks to the health and safety or fundamental rights of persons, which some deep fakes definitely do, including causing psychological harm to groups and individuals [77]. Deep fakes can also benefit from subliminal techniques [78] that are generally prohibited pursuant to Article 5(1a) of the AI Act. Problems may arise, however, in qualifying at which point a deep fake becomes a subliminal deep fake. Such difficulties may occur in the case of microtargeted video deep fakes based on facial resemblance and mimicry, which increase trustworthiness or self-enhancement among recipients [79–81].

The pillar on which the system the AI Act is built on is trustworthiness. Unfortunately, this system has some gaps as it mainly concentrates on the intended uses of specific AI systems and 'applies mandatory requirements for pre-defined domains of use', leaving some misuses and abuses unregulated [82]. Leaving deep fakes outside the scope of the high-risk category matches the general concept behind the AI Act risk assessment, but it does not take into account the fundamental malicious misuses of technology.

In the author's opinion, the reasonable solution for now would be to leave deep fakes within the low or minimal risk category with specific transparency obligations and distinguishing very concrete subgroups/exceptions for reclassification into the high-risk category or even imposing direct bans, which is needed in the case of deep porn [16]. Another issue is the fundamental effectiveness of the permissions, bans and transparency obligations, which will be discussed later.

From the point of view of strengthening social awareness and resilience, it is important to indicate why deep fakes give rise to threats – the AI Act might be the most appropriate place for the proper remarks. The European Parliament's Committee on the Internal Market and Consumer Protection and the European Parliament's Committee on Civil Liberties, Justice and Home Affairs [83] were clearly not sure about the qualification of deep fakes as a specific risk AI system. In a draft report from April 2022, it was proposed to add Recital 40a to the AI Act. It was supposed to clearly state:

> *Certain AI systems should at the same time be subject to transparency requirements and be classified as high-risk AI systems, given their potential to deceive and cause both individual and societal harm. In particular, AI systems that generate deep fakes representing existing persons have the potential to both manipulate the natural persons that are exposed to those deep fakes and harm the persons they are representing or misrepresenting, while AI systems that, based on limited human input, generate complex text such as news articles, opinion articles, novels, scripts, and scientific articles ('AI authors') have the potential to manipulate, deceive, or to expose natural persons to built-in biases or inaccuracies.*

This is a particularly interesting approach, questioning the Commission's initial qualification. Attention was rightly paid to the potential of deep fakes to deceive or cause harm to individuals and society. Such a comment – regardless of the final qualification of deep fakes – should appear within the AI Act to highlight the problem of manipulation, as well as the huge, often irreparable damages inflicted on individuals [84].

Interestingly, the German Bundesrat [85] was one of the few European chambers of parliaments to refer to the Commission's proposal in a resolution from September 2021 and touch upon the issue of deep fakes directly. It was rightly emphasised that deep fakes can manipulate public discourse in a covert manner, thereby exerting a significant influence on the process of individual and public opinion formation, and that they should not be treated as a side effect of the use of AI. It was suggested to consider deep fakes as a high-risk AI system and foreseen that this part of media law would have to be addressed properly by Member States since the AI Act does not cover that dimension properly [85].

It cannot be ruled out that in future deep fakes will become the subject of further thorough analyses and will be included in the list of high-risk AI systems. This type of evaluation will have to take into account, above all, the development of technology and its actual applications, for which permanent case study monitoring is necessary. First of all, it will be necessary to evaluate the validity and effectiveness of the introduced countermeasures. There is a high probability that transparency obligations alone will be insufficient to stop the vast majority of deep fakes of a malicious nature and that even moving to a higher risk category and becoming subject to strict obligations will not significantly change these negative trends. The remark from the Bundesrat in regard to the engagement of Member

States may actually indicate what path to combat deep fakes will become a future priority.

4. Transparency obligations and disclosure rules

The European Parliament's Committee on Legal Affairs [86], already at the beginning of 2021, indicated that deep fakes should be generally covered by disclosure rules, as they *could be used to blackmail, generate fake news reports, or erode public trust and influence public discourse; (...) such practices have the potential to destabilise countries, spreading disinformation and influencing elections.* The AI Act followed up on that assumption, though the regulation itself does not directly refer to the above-mentioned misuses of deep fakes.

Initially [4], it was proposed by the Commission that deep fakes would be classified as systems for which 'minimum transparency rules' would be required. This approach aroused justified controversy due to the threats associated with the presence of deep fakes in the information space. Mesarčík et al. [22] rightly pointed out that the proposed obligations lacked robustness and did not have the potential to significantly 'reduce the information asymmetry and thus allow the users (citizens) to combat the effects of deepfakes and still form informed and accurate opinions'.

The key to regulating the transparency obligations for deep fakes is Article 52(3) of the AI Act, which was fundamentally extended by the European Parliament. Initially [4], it contained an extremely general provision: *shall disclose that the content has been artificially generated or manipulated.* The amended version of Article 52(3) [24] introduces much more specific regulations that allow us to look at the solutions with cautious optimism:

> *shall disclose in an appropriate, timely, clear and visible manner that the content has been artificially generated or manipulated, as well as, whenever possible, the name of the natural or legal person that generated or manipulated it. Disclosure shall mean labelling the content in a way that informs that the content is inauthentic and that is clearly visible for the recipient of that content. To label the content, users shall take into account the generally acknowledged state of the art and relevant harmonised standards and specifications.*

Additionally, the European Parliament [24] has rightly added Article 52(3b), addressing some features of disclosure, and introduced special protection for vulnerable persons:

> *The information referred to in paragraphs 1 to 3 shall be provided to the natural persons at the latest at the time of the first interaction or exposure. It shall be accessible to vulnerable persons, such as persons with disabilities or children, complete, where relevant and appropriate, with intervention or flagging procedures for the exposed natural person taking into account the generally acknowledged state of the art and relevant harmonised standards and common specifications.*

This is definitely a step in the right direction and another of the significant and positive changes to the draft of the AI Act proposed by the European Parliament. The phrases 'appropriate, timely, clear and visible' seem to be of extreme importance, but it should be remembered that only standardisation processes allowing for the introduction of clear disclosure rules will enable final assessment of the adopted solutions and measuring their effectiveness in regard to some deep fakes (those that will be subject to any transparency obligations at all).

The Commission did not specify who would be the addressee of the disclosure [34]. The Parliament's amendments are more precise in this regard, even if they refer to the broad term of 'recipients'. In regard to deep fakes, transparency obligations are primarily meant to sensitise recipients and raise their awareness, or even serve to protect 'some right to reality grounded in fundamental rights' [23]. They are intended to show that recipients are dealing with fake content that does not represent reality – either distorting it in its entirety or falsifying it in order to mislead the audience [21]. The Explanatory Memorandum of the AI Act indicates that the obligation to disclose should allow recipients to make informed choices or step back from a given situation. The early-warning system is aimed at protecting recipients, their awareness and, to a large extent, trust in the information system. The positive impact of disclosure rules should then be considered mainly in the context of disinformation or media consistency. The rationale behind the provisions seems to be clear – deep fakes must be properly labelled due to their deceptive potential.

However, it should be considered whether transparency obligations will actually effectively protect recipients against disinformation. Expecting state or non-state actors with malicious goals to comply

with AI disclosure rules is obviously irrational. Rather, it should be assumed that transparency obligations will play a role in reducing the number of deep fakes circulating in the information space, especially those created by users equipped with unsophisticated software, but will not be a barrier for specialised actors.

In 2023 alone, deep fakes were successfully used in the US, Turkey and Germany, where they played a role in either influencing the election results or in fuelling current divisive issues. In Turkey, one of the opposition candidates in the presidential elections, Muharrem İnce, fell victim to deep porn and had to withdraw his candidature. İnce accused Russia of meddling in the Turkish elections [87]. In 2022, Russia used a deep fake video depicting the President of Ukraine, Volodymyr Zelensky, who was supposedly calling on his troops to surrender [88]. In the US, supporters of the Republican Party's rival candidates – Donald Trump and Ron DeSantis – continuously publish deep fake images and videos ridiculing their opponents [89]. US President Joe Biden is regularly the target of falsified information intended to damage his reputation, especially in the context of the 2024 US elections [90]. In Germany, deep fake videos depicting Canadian psychologist Jordan Peterson were disseminated to discredit the Minister of Foreign Affairs, Annalena Baerbock [91], while Minister of Economy Robert Habeck allegedly announced the closure of all outdoor swimming pools in response to incidents of violence [92]. The latter incident was intended to cause additional social unrest.

In 2023, a disturbing trend of using images of public figures to publish hate speech, anti-Semitic, racist or misogynistic content was observed. The voice of the popular actress Emma Watson was used to generate an audio deep fake in which she read fragments of Adolf Hitler's 'Mein Kampf' [93]. Journalists Joe Rogan and Ben Shapiro allegedly made homophobic and transphobic remarks [94].

Although most manipulations seem to be internally driven, the influence of external actors, including foreign countries, in cases of a strictly political nature, cannot be ruled out. The outreach and impact of content is generally multiplied by public willingness to share it, which mirrors the patterns of spreading disinformation due to injecting 'false but compelling information into a ready and willing information-sharing environment' by ordinary users [6].

In fact, transparency obligations in the form introduced by the AI Act could not be enforced in most of the cases mentioned above (assuming the applicability of the law due to jurisdiction). This results

directly from the intentions of its authors, which include, first and foremost, intentional and conscious misleading of recipients. As some deep fakes are created for the purposes of foreign information manipulation and interference, it should be assumed that the state and non-state actors involved in this practice will, for obvious reasons, not comply with any transparency obligations. In this context, simple technical solutions based on disclosure will be toothless [75, 95]. Therefore, the solutions proposed in the AI Act do not fully take into account the specificity of creating and disseminating deep fakes, the context of international politics and already known patterns of disinformation.

As a result, the transparency obligations 'will be applicable to only a small portion of deep fakes' [75]. The analysis by M. Veale and F. Z. Borgesius [23], who are quite critical of the way deep fakes were regulated in the AI Act, rightly pointed out that 'disclosure may only partially assist the subject', which in view of potential limitations on the effects of disclosure rules, may not be sufficient.

Protected goods must also include the personal rights of third parties whose image is the subject of the synthesis. Unfortunately, disclosure alone would not protect the subject/object of the depiction entirely. The organisation Access Now rightly pointed out that in many cases the 'transparency obligation will be insufficient to mitigate all risks associated with such applications' [96]. It needs to be clearly stated that in regard to deep porn, transparency obligations would not prevent the victimisation of depicted persons [75]. It is similar in cases of defamation or discrediting of individuals, when deep fakes can act as a catalyst for long-term negative emotions and associations. Research on the long-term consequences of exposure to fake news has shown that prior exposure increases the perceived accuracy of fake news [97]. Disclosure would not be able to stop these processes entirely. It would also not counteract the negative phenomenon of increased uncertainty in the case of exposure to fake content that might in turn undermine trust in the media, as proven by the experiment conducted by C. Vaccari and A. Chadwick [33]. The connection between disclosure and the actual reactions of recipients to AI-generated content could become the subject of research involving an evaluation of neural pathways and the possible outcomes of interference between two different messages – false information and disclosure of the falsehood.

It should also be noted that the AI Act imposes transparency obligations on 'users', while in the case of chatbots, it refers to 'providers'. Also in this regard, one may have concerns as to whether the transfer

of the burden to users is justified [20, 23], or even if provisions might be creating a 'legal loophole' [98]. A similar point was made by N. Helberger and N. Diakopoulos [68], who indicated that responsibility for the use of AI systems should lie primarily with providers, not users. The opposite would shift responsibility to end-users and disregard the potential risks of misusing certain systems.

Expanding the scope of provisions in regard to deep fakes might extend to potential additional legal obligations for providers [16]. EU regulations might oblige software providers to comply with fundamental rights and require further transparency [99], which in turn would add certainty to introduced solutions [98]. Moreover, the EU should understand the weaknesses of the AI Act in relation to counteracting deep fakes in order to consistently increase the legal regime in other areas. The postulated synergy effect between the AI Act and the DSA or Strengthened Code of Practice on Disinformation [21] seems to be a rational approach that takes into account various aspects of the negative impact of deep fakes. It should be emphasised again that precision, internal consistency and solution complementarity are necessary in this respect.

Another problem seems to be label parameters. Undoubtedly, markings should appear at the beginning of the material (pursuant to Article 52 point 3b of the AI Act [24], the information *shall be provided to the natural persons at the latest at the time of the first interaction or exposure*), though technical solutions might vary depending on the form of media used. In the case of video deep fakes, it seems advisable to disclose the fake nature of the content throughout playback in text form, so recipients are constantly aware that their interaction is based on AI activity. In the case of image deep fakes, it should be clearly and visibly stated in text form and disclosure should be an integral part of the image. In the case of audio deep fakes, it seems advisable to adjust disclosure to the length of the audio and the information should be read at least at the beginning and at the end of the display. Standardisation processes should take into account existing regulations and experiments on forms of disclosure. One interesting example is Bill S.B. 5152, adopted by the Washington State Senate in 2023[6] [100].

One might have reservations about the form of disclosure if only fragments of audio or visual content have been manipulated. It seems reasonable to ask whether, as a rule, the general pattern of AI disclosure should apply, or whether it should be modifiable and indicate which part of the material bears traces of AI interference [34]. In the author's opinion, it seems reasonable to label entire

6 —— Washington State Senate Bill on Defining synthetic media in campaigns for elective office, and providing relief for candidates and campaigns (S.B. 5152) [100] states: (4) It is an affirmative defense for any action brought under this section that the electioneering communication containing a synthetic media includes a disclosure stating, 'This (image/video/audio) has been manipulated,' in the following manner: a) For visual media, the text of the disclosure must appear in a size easily readable by the average viewer and no smaller than the largest font size of other text appearing in the visual media. If the visual media does not include any other text, the disclosure must appear in a size that is easily readable by the average viewer. For visual media that is a video, the disclosure must appear for the duration of the video; or (b) If the media consists of audio only, the disclosure must be read in a clearly spoken manner and in a pitch that can be easily heard by the average listener, at the beginning of the audio, at the end of the audio, and, if the audio is greater than two minutes in length, interspersed within the audio at intervals of not more than two minutes each.

content without any distinctions to avoid further manipulation and misleading of recipients.

The importance of appropriate labelling is emphasised by persistent low social awareness. Research conducted by Bitkom in Germany [102] shows that only 15% of respondents are able to explain what a deep fake is, and a mere 23% have basic knowledge on the subject. As many as 84% of respondents are in favour of marking deep fakes. Taking into account the very low number of respondents who are familiar with deep fakes, labelling must be adjusted to different kinds of audiences, which might be partly achieved by using simplified, concrete language.

A standardisation effort will be necessary in this regard. The Commission 'has begun to adopt a standardisation request which will provide a formal mandate to European standardisation organisations to develop standards under the AIA [AI Act]' [103]. Similarly, Article 82b(1) of the AI Act added by the European Parliament [24] indicates that the Commission shall develop, in consultation with the AI office, guidelines on the practical implementation of this Regulation, and in particular on the practical implementation of transparency obligations laid down in Article 52.

This area of research seems to be understudied and researchers need to enhance the outcome of standardisation processes. However, a very recent research study conducted by Dutch scientists dealing with deep fakes [104] is noteworthy, as it simulated the marking of video materials using three colours: green (veracity confirmed), yellow (veracity not confirmed), red (content containing false messages). The research results show that even such basic disclosure significantly increases scepticism among recipients and affects credibility assessments of the material. The researchers also tested the display time of the messages. Undoubtedly, such experiments must be repeated and modified in future to work out the best possible formula to measure when exactly labels should be displayed.

One could plausibly argue that even disclosure would not solve the problem of vulnerability to manipulation, or that the correlation between mere disclosure of using an AI system and increased protection of fundamental rights is relatively weak [55], but disclosure alone is a first step to protection and reduction of the negative effects of (some) deep fakes. An additional solution might be watermarking deep fake content [105], authenticating real content, or strengthening cyberliteracy to raise awareness among recipients.

5. Exceptions

Transparency obligations for deep fakes provide certain exceptions to the basic principles. The European Parliament [24] has made some significant changes to the Commission's proposal, also by extending and specifying the scope of exceptions. After amendment, Article 52(3a) states:

> *Paragraph 3* [transparency obligations] *shall not apply where the use of an AI system that generates or manipulates text, audio or visual content is authorised by law or if it is necessary for the exercise of the right to freedom of expression and the right to freedom of the arts and sciences guaranteed in the Charter of Fundamental Rights of the EU, and subject to appropriate safeguards for the rights and freedoms of third parties. Where the content forms part of an evidently creative, satirical, artistic or fictional cinematographic, video games visuals and analogous work or programme, transparency obligations set out in paragraph 3 are limited to disclosing of the existence of such generated or manipulated content in an appropriate clear and visible manner that does not hamper the display of the work and disclosing the applicable copyrights, where relevant. It shall also not prevent law enforcement authorities from using AI systems intended to detect deep fakes and prevent, investigate and prosecute criminal offences linked with their use.*

The exceptions therefore include two basic groups:

- authorisation by law (and in a later part – detection of deep fakes);
- exercise of the right to freedom of expression and the right to freedom of the arts and sciences that includes evidently creative, satirical, artistic or fictional cinematographic, video games visuals and analogous work or programme.

The former point does not seem to be controversial. It is the latter that has the potential to cause interpretation problems. Hertie School of Governance experts [75] predict that exceptions will open the door to creative manipulation and are 'likely to bring inconsistencies in practice'. The proposal rightly seeks to ensure high-level protection of the fundamental rights guaranteed by the EU Charter of Fundamental Rights, including freedom of expression (Article 11), or freedom of art and science (Article 13), but the practice may prove treacherous as the system of exceptions would possibly pave the way for exploitation.

Omission of the term 'timely' in Article 52(3a) of the AI Act, in comparison to Article 52(3) [24], leads to unnecessary problems of interpretation, especially since the legislator's intention was apparently to approximate the provisions on non-exceptions and exceptions within this particular AI system. Based on very general formulations, it is difficult to determine what disclosure would actually look like in the case of exceptions. The distinction itself opens up room for manipulation and misinterpretation.

In an increasing number of cases, legislators have prohibited or limited the use of deep fakes, but they have also allowed significant exceptions in the form of obvious or evident satire or parody (of a 'demonstrably' fake nature) [25, 106–108]. This 'obvious' or 'evident' nature may be debatable and would have to be assessed on a case-by-case basis because it might depend on contextualisation as well as the cognitive abilities, media knowledge, or social and political awareness of recipients.

Unfortunately, overusing the legal exemptions could be seen as a useful tool to circumvent the restrictions. Deep fakes are described as a phenomenon that might benefit from the 'just joking' excuse, making it possible to smuggle illegal content or manipulate the audience 'under the guise of humour', which might even lead to the 'weaponisation of humour' [109]. Satirical context has already been shown to function as 'a cover for spreading' extremist ideologies [25] with respect to fake news. At the same time, the fight against deep fakes might also be used to justify suppressing freedom of speech. This is especially important in the case of non-democratic countries that hide their censorship tendencies under the guise of protecting social stability [15].

While deep fakes can be successfully used to create content that is critical of the authorities, the limits of satire are hard to grasp, especially since the boundaries between satire and harmful content are increasingly blurred. Difficulties also arise when 'satire is transferred out of its original context', is 'no longer recognisable' due to high synthesis quality, or is not recognised by recipients [110].

One potential solution could be to treat all deep fakes in the same way with respect to transparency obligations. If the satirical or parodic nature of the material is obvious, disclosing the use of AI and appropriately flagging the fake content should not be a problem and standardised transparency obligations would help to protect recipients. It seems reasonable to refer to fundamental rights, including freedom of speech, while noting that the requirements regarding

transparency obligations will not violate these rights. Labelling AI-manipulated audio or visual content should be seen as a standard rather than an arduous obligation.

6. Conclusions

Ongoing work on the AI Act in regard to deep fakes gives hope for more robust protection of EU citizens against AI manipulation. It should be emphasised that the amendments introduced by the European Parliament would slightly increase the ability of the EU to counteract the negative effects of deep fakes. The imperfection of the solutions results to a large extent from general legislative difficulties related to the creation and dissemination of deep fakes, the specificity of deep fakes and the complexity of the challenges they create for democratic systems, societies and individuals, but it is also due to an internal lack of EU coherence or precision.

The European Parliament has already introduced numerous positive changes to the Commission's proposal for the AI Act, thus addressing some of the critical analyses by experts. However, this does not mean that the regulation is free of deficits in its current form. A detailed analysis of the Commission's proposal and European Parliament amendments in regard to deep fakes allows us to draw a number of conclusions and identify reservations that should be considered in future revisions of the AI Act.

1. The level of expectations should be adjusted to the AI Act's true capabilities to influence reality, especially since the provisions on deep fakes are not the key element of the regulation and the level of protection it offers against them is basic at best. The authors of a significant portion of deep fakes will neither comply with transparency obligations nor care about the risk categories [23, 75]. As emphasised, the vast majority of deep fakes comprises non-consensual pornography (deep porn). Such materials, due to their specific nature and manner of dissemination, will never be subject to any disclosure rules. In this context, it is necessary to implement stricter provisions aimed at protecting individuals, in particular women, against the deployment of gender-based violence, exploitation, humiliation, or harassment. The European Parliament resolution, containing recommendations to the Commission on combating gender-based violence [54], paves the way for further actions. This might be achieved through additional countermeasures, including putting pressure on platforms that enable the dissemination of such content,

which the DSA fortunately already does. In the case of most non-consensual malicious deep fakes (including deep porn), the basic idea of creating and disseminating deep fakes violates the law, even if the relevant provisions are derived from civil, tort, criminal or copyright law. Member States have to reconsider how to make these provisions more efficient. Therefore, the role that the AI Act would play in combating deep porn would be significantly reduced and one should not expect a breakthrough. The proposed transparency obligations seem to be appropriate to regulate a small portion of deep fakes appearing in the information space. This applies not only to deep porn but also to some disinformation activities that might be driven by foreign information manipulation and interference.

2. Hertie School of Governance experts [75] rightly pointed out that the AI Act offers the 'false promise of transparent deep fakes'. Disclosure rules give the illusory belief that revealing the false nature of content (if it gets done at all) will lead to the elimination of the negative effects of creating and disseminating deep fakes. It will not. The problem with deep porn or discrediting materials is the non-consensual use of someone else's image and the psychological and reputational harm it creates [26]. Even if disclosure rules are applied to non-consensual deep fakes (especially deep porn), the negative effects leading to psychological harm will not be eliminated. Many women have been victims of non-consensual pornography and have reported severe psychological effects, including discomfort while using social media, depression, anxiety or trauma [26, 96, 111]. Similar consequences can be measured with respect to false content of a discrediting nature since malicious deep fakes can cause reputational harm and thus have long-lasting repercussions on the psychological well-being or professional prospects of the depicted individuals [112].

3. The EU must ensure internal coherence, in particular in regard to the proposed definitions and descriptions of deep fakes. Therefore, there should be absolutely no internal discrepancies within the AI Act or between different legal acts proposed by the EU. The certainty of the law, its interpretation and enforcement must be an asset of EU legislative activity. The AI Act may set a common standard, to which subsequent legal acts will refer. Deep fakes must be unambiguously defined, and the definition must clearly include, among others, the scope of the form of deep fakes (typological aspect) and the subjects/objects to which deep fakes refer (subjective aspect). That applies mainly

to discrepancies between Recital 70, Article 3(1) point 44d and Article 52(3) of the AI Act. In the author's opinion, it is necessary to extend the scope of the definition to 'persons, objects, places or other entities or events', as well as to reconsider the potential omission of deep fakes in textual form.

4. Transparency obligations are not a universal solution. J. Habgood-Coote [95] may be right in pointing out that a significant number of researchers is guided by 'technochauvinism' or 'techno-fixation', assuming that the problem of deep fakes can be solved with the use of technological tools. It might be better to qualify deep fakes 'as a social problem about the management of our practices for producing and receiving recordings' [95]. At the same time, technological solutions can at least reduce negative trends, acting as a deterrent. That is why it is so important to find a balance between various ways of counteracting the harmful uses of deep fakes and their negative consequences, which might include disclosure, watermarking, content authentication, or strengthening cyberliteracy [12]. Even if it would help to eliminate only a small number of deep fakes, the AI Act should be seen as a step in the right direction, but it needs to be supplemented with further regulatory and non-regulatory efforts from the EU to strengthen social resilience, also by enhancing cyberliteracy. Again, omitting the critical reference to the specific risks that deep fakes pose overlooks a significant aspect of raising awareness through the AI Act.

5. If transparency and disclosure are to introduce a reasonable level of protection, it is necessary to tighten the system to prevent possible attempts to circumvent the obligations. It is advisable to reconsider sealing the system of exceptions to full disclosure rules. The assumption that the satirical or parody nature of the material is 'obvious' or 'evident' is based on a misconception about the high level of cognitive and analytical skills[7] among recipients of deep fakes [110]. Satire and parody can be successfully used to bypass some safeguards in order to smuggle sophisticated political manipulation and thus influence the audience. It will also be crucial to develop the practice for disclosure rules, which requires standardisation processes and empirical research to measure the effectiveness of different solutions. Ongoing work by Dobber et al. [104] as well as solutions introduced in the US might serve as an example. The EU should closely monitor regulatory efforts in other countries to either use the labelling patterns for standardisation processes, or even introduce concrete provisions within the AI Act. The transparency

7 — J. Langa [108], commenting on the provisions introduced in the US, refers to the notion of 'reasonable person' that 'realizes that a deepfake is satirical or parodical' and thus cannot be deceived. The term seems to be vague and the highly deceptive nature of deep fakes (especially video deep fakes) has been proven many times.

obligations might be complemented by imposing additional obligations on providers and manufacturers. Although this will not eliminate non-consensual deep fakes of a harmful nature, it will limit their effects and the amount of manipulated and unmarked content by making it more difficult for non-specialised users to create deep fakes [98].

6. The AI Act does not impose any special obligations on digital platforms in regard to the creation and dissemination of deep fakes. It can be argued that such solutions are found in other acts introduced by the EU, but the lack of an internal connection does not directly indicate the specific purpose and complementarity of measures counteracting deep fakes [75]. According to some experts [12], 'distribution and consumption patterns pose larger threats to democracy and society than the fake content itself'. It might be advisable to concentrate on prevention by delimiting the applications of technology, also through ethical norms [113], and reducing dissemination capabilities. The Centre for Data Innovations [114] suggested 'nimbler soft law approaches' to 'supplement adjustments to the AI Act and the Directive on Gender-based Violence' by working closely with industry and encouraging self-regulatory efforts to counteract non-consensual pornography. That would definitely fit with the idea of reducing the impact of deep fakes by restricting amplification of the content through online platforms [76], which applies not only to deep porn but also to other types of deep fakes, including those of an intrinsically political nature.

7. The European Parliament's amendments in regard to deep fake detection systems in the form of deletion from the list of high-risk AI systems should be assessed positively, primarily due to an initial erroneous discrepancy in the risk assessment between deep fakes and the technological measures that are intended to protect against them. The potential problem with the provision included in paragraph 1 point 6d in Annex III of the AI Act, which might lead to the indirect inclusion of deep fake detection systems in the high-risk AI systems list [70], might pose interpretational problems and should be clarified at a later stage in the negotiations.

8. It still seems controversial that deep fakes are not qualified to the category of high-risk AI systems, especially because the AI Act provides some rationale for reclassification. The potential solution might be to single out those deep fakes that pose

a greater threat to specific subjects for special protection and transfer them to a higher category, describing the scope of their harmfulness in a clear and precise manner to leave no room for misinterpretation (e.g. introducing additional protection for candidates before an election), or introducing a complete ban on their creation and dissemination (e.g. deep porn). The omission of the contextual aspect while assessing the risk posed by deep fakes can be assessed negatively. Reference to the harmful uses of deep fakes and detrimental effects they cause should be added in one of the recitals, which might also be extended by broader and more-detailed reference to the systemic and societal harms that AI systems might pose [65, 115]. At the same time, it should be remembered that even moving to a higher risk category will not be a universal solution or eliminate the basic problem related to the spread of some deep fakes of a malicious nature, since they are not subject to any rules.

9. The fundamental problem with the emergence of deep fakes in the information space is not a complete lack of regulation. In many cases, deep fakes of a malicious nature are directly or indirectly prohibited by law, and victims can pursue their rights in court. The problem, however, is enforcement of existing provisions [59]. The AI Act would not change this situation drastically, and some may rightly accuse the regulation of failing to impose sanctions for non-compliance with the transparency obligations [74]. These can easily be derived from other EU legal acts, including the DSA, which obliges platforms to inform users about the deceptive or manipulative nature of content [116]. Pursuant to the DSA, non-compliance can be sanctioned by up to 6% of annual worldwide turnover. The AI Act should complement these solutions, particularly with respect to authors of deep fakes and AI system providers [21], which currently is not the case. Although numerous researchers have pointed out that identifying perpetrators is problematic (the basic problem with attribution), the AI Act might add another source of pressure and mobilise law enforcement authorities and policymakers to deal with the problem [25, 98].

References

[1] J.Laux, S. Wachter, B. Mittelstadt, "Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk," *SSRN Electronic Journal*, 2023, doi: 10.2139/ssrn.4230294.

[2] P. Maham, S. Küspert, *Governing General Purpose AI.* Stiftung Neue Verantwortung, Berlin, 2023.

[3] J. Schuett, "Risk Management in the Artificial Intelligence Act," *European Journal of Risk Regulation*, pp. 1–19, 2023, doi:10.1017/err.2023.1.

[4] European Commission, *Proposal for Regulation of the European Parliament and the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts.* COM(2021) 206 final. 2021. [Online]. Available: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF. [Accessed: Sep. 13, 2023].

[5] T. Brooks, G. Princess, J. Heatley, J. Jeremy, K. Scott, et al., *Increasing Threats of Deepfake Identities,* U.S. Department of Homeland Security, 2019. [Online]. Available: https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf. [Accessed: Sep. 13, 2023].

[6] B. Chesney, D. Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review*, vol. 107, no. 18, pp. 1753–1820, 2019, doi: 10.15779/Z38RV0D15J

[7] I. Dąbrowska, "Deepfake – nowy wymiar internetowej manipulacji," *Zarządzanie Mediami*, vol. 8, no. 2, pp. 89–101, 2020, doi: 10.4467/23540214ZM.20.024.11803.

[8] D. L. Byman, C. Gao, C. Meserole, *Deepfakes and international conflict*. Washington: The Brookings Institution, 2023.

[9] M. Pawelec, "Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions," *Digital Society*, vol. 2, no. 2, 2022, doi: 10.1007/s44206-022-00010-6.

[10] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, Y. K. Dwivedi, "Deepfakes: Deceptions, mitigations, and opportunities," *Journal of Business Research*, vol. 154, 2023, doi: 10.1016/j.jbusres.2022.113368.

[11] E. Pashentsev, "The Malicious Use of Deepfakes Against Psychological Security and Political Stability," in *The Palgrave Handbook of Malicious Use of AI and Psychological Security*, E. Pashentsev, Ed. London: Palgrave Macmillan, Cham, 2023, pp. 47–80.

[12] H. Farid, H.-J. Schindler, *Deep Fakes. On the Threat of Deep Fakes to Democracy and Society*. Berlin: Konrad Adenauer Stiftung, 2020.

[13] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, et al., "Countering Malicious DeepFakes: Survey, Battleground, and Horizon," *International Journal of Computer Vision*, vol. 130, no. 7, pp. 1678–1734, 2022, doi: 10.1007/s11263-022-01606-8.

[14] J. Ice, "Defamatory Political Deepfakes and the First Amendment," *Case Western Reserve Law Review*, vol. 70, no. 2, pp. 417–455, 2019.

[15] E. Hine, L. Floridi, "New deepfake regulations in China are a tool for social stability, but at what cost?," *Nature Machine Intelligence*, vol. 4, no. 7, pp. 608–610, 2022.,doi: 10.1038/s42256-022-00513-4.

[16] R. T. Toparlak, "Criminalising Pornographic Deep Fakes: A Gender-Specific Inspection of Image-Based Sexual Abuse," *SciencesPo Law School The 10th Graduate Conference*, 2022. [Online]. Available: https://www.sciencespo.fr/public/chaire-numerique/wp-content/uploads/2022/06/3a-Toparlak_Criminalising-Pornographic-Deep-Fakes.pdf. [Accessed: Sep. 13, 2023].

[17] E. Meskys, A. Liaudanskass, J. Kalpokiene, P. Jurcy, "Regulating Deep-Fakes: Legal and Ethical Considerations," *Journal of Intellectual Property Law & Practice*, vol. 15, no. 1, pp. 24–31, 2020, doi: 10.1093/jiplp/jpz167.

[18] K. Mamak, "Categories of Fake News from the Perspective of Social Harmfulness," in *Integrity of Scientific Research: Fraud, Misconduct and Fake News in the Academic, Medical and Social Environment*, J. Faintuch, S. Faintuch, Ed. Springer, pp. 351–357, 2022, doi: 10.1007/978-3-030-99680-2_35.

[19] M. Kop, "EU Artificial Intelligence Act: The European Approach to AI," *Transatlantic Antitrust and IPR Developments*, vol. 2, 2021. [Online]. Available: https://law.stanford.edu/wp-content/uploads/2021/09/2021-09-28-EU-Artificial-Intelligence-Act-The-European-Approach-to-AI.pdf. [Accessed: Sep. 13, 2023].

[20] L. Edwards, *The EU AI Act: a summary of its significance and scope*, Ada Lovelace Institute, 2022. [Online]. Available: https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf. [Accessed: Sep. 13, 2023].

[21] A. Fernandez, "'Deep fakes': disentangling terms in the proposed EU Artificial Intelligence Act," *UFITA Archiv für Medienrecht und Medienwissenschaft*, vol. 85, no. 2, pp. 392–433, 2021, doi: 10.5771/2568-9185-2021-2-392.

[22] M. Mesarčík, S. Solarova, J. Podroužek, M. Bielikova, *Stance on The Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence – Artificial Intelligence Act*, Kempelen Institute of Intelligent Technologies, 2021, doi: 10.31235/osf.io/yzfg8.

[23] M. Veale, F. Z. Borgesius, "Demystifying the Draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach," *Computer Law Review International*, vol. 22, no. 4, pp. 97–112, 2021, doi: 10.9785/cri-2021-220402.

[24] European Parliament, *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.* (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))1. P9_TA(2023)0236. 2023. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf. [Accessed: Sep. 13, 2023].

[25] M. van Huijstee, P. van Boheemen, D. Das, L. Nierling, J. Jahnel et al., *Tackling deep-fakes in European policy*, European Parliamentary Research Service, Brussels, 2021.

[26] R. Delfino, "Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act," *Fordham Law Review*, vol. 88, no. 3, pp. 887–938, 2019, doi: 10.2139/ssrn.3341593.

[27] A. De Ruiter, "The Distinct Wrong of Deepfakes," *Philosophy & Technolology*, vol. 34, pp. 1311–1332, 2021, doi: 10.1007/s13347-021-00459-2.

[28] H. Farid, "Creating, Using, Misusing, and Detecting Deep Fakes," *Journal of Online Trust and Safety*, vol. 1, no. 4, pp. 1–33, 2022, doi: 10.54501/jots.v1i4.56.

[29] A. Satariano, P. Mozur, *The People Onscreen Are Fake. The Disinformation Is Real,* 2023. [Online]. Available: https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html. [Accessed: Sep. 13, 2023].

[30] T. Weikmann, S. Lecheler, "Cutting through the Hype: Understanding the Implications of Deepfakes for the Fact-Checking Actor-Network," *Digital Journalism*, 2023, doi: 10.1080/21670811.2023.2194665.

[31] M. Westerlund, "The Emergence of Deepfake Technology: A Review," *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39–52, 2019, doi: 10.22215/timreview/1282.

[32] S. J. Nightingale, H. Farid, "AI-synthesized faces are indistinguishable from real faces and more trustworthy," *Proceedings of the National Academy of Sciences,* vol. 119, no. 8, 2022, doi: 10.1073/pnas.2120481119.

[33] C. Vaccari, A. Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," *Social Media + Society*, vol. 6, no. 1, 2020, doi: 10.1177/2056305120903408.

[34] B. van der Sloot, Y. Wagensveld, "Deepfakes: regulatory challenges for the synthetic society," *Computer Law & Security Review*, vol. 46, 2022, doi: 10.1016/j.clsr.2022.105716.

[35] C. Okolie, "Artificial Intelligence-Altered Videos (Deepfakes), Image-Based Sexual Abuse, and Data Privacy Concerns," *Journal of International Women's Studies*, vol. 25, no. 2, 2023. [Online] Available: https://vc.bridgew.edu/jiws/vol25/iss2/11. [Accessed: Sep. 13, 2023].

[36] L. Hurst, *How a fake image of a Pentagon explosion shared on Twitter caused a real dip on Wall Street*, 2023. [Online]. Available: https://www.euronews.com/next/2023/05/23/fake-news-about-an-explosion-at-the-pentagon-spreads-on-verified-accounts-on-twitter. [Accessed: Sep. 13, 2023].

[37] C. Öhman, "The identification game: deepfakes and the epistemic limits of identity," Synthese, vol. 200, no. 4, 2022, doi: 10.1007/s11229-022-03798-5.

[38] A. Kleemann, "Deepfakes – Wenn wir unseren Augen und Ohren nicht mehr trauen können," *SWP-Aktuell*, vol. 43, 2023, doi: 10.18449/2023A43.

[39] K. Eckart, *A growing problem of 'deepfake geography': How AI falsifies satellite images*, 2021. [Online]. Available: https://www.washington.edu/news/2021/04/21/a-growing-problem-of-deepfake-geography-how-ai-falsifies-satellite-images. [Accessed: Sep. 13, 2023].

[40] B. Zhao, S. Zhang, C. Xu, Y. Sun, C. Deng, "Deep fake geography? When geospatial data encounter Artificial Intelligence," *Cartography and Geographic Information Science*, vol. 48, no. 4, pp. 338–352, 2021, doi: 10.1080/15230406.2021.1910075.

[41] K. Hiebert, Democracies Are Dangerously Unprepared for Deepfakes, 2022. [Online]. Available: https://www.cigionline.org/articles/democracies-are-dangerously-unprepared-for-deepfakes. [Accessed: Sep. 13, 2023].

[42] European Parliament's Committee on Industry, Research and Energy, *Opinion on the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts,* 2022. [Online]. Available: https://www.europarl.europa.eu/doceo/document/ITRE-AD-719801_EN.pdf. [Accessed: Sep. 13, 2023].

[43] J. Bateman, *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*. Washington: Carnegie Endowment for International Peace, 2020.

[44] K. Giles, K. Hartmann, M. Mustaffa, *The Role of Deepfakes in Malign Influence Campaigns*. Riga: NATO Strategic Communications Centre of Excellence, 2019.

[45] J. R. Allen, D. M. West (2020). *The Brookings glossary of AI and emerging technologies*. [Online]. Available: https://www.brookings.edu/articles/the-brookings-glossary-of-ai-and-emerging-technologies. [Accessed: Sep. 13, 2023].

[46] T. Hwang, *Deepfakes: A Grounded Threat Assessment*. Washington: Center for Security and Emerging Technology, 2020.

[47] R. Mattioli, A. Malatras, *Identifying Emerging Cyber Security Threats and Challenges for 2030.* Athens: ENISA, 2023.

[48] M. Atleson, *Combatting Online Harms Through Innovation.* Federal Trade Commission Report to Congress, 2022. [Online] Available: https://www.ftc.gov/system/files/ftc_gov/pdf/Combatting%20Online%20Harms%20Through%20Innovation%3B%20Federal%20Trade%20Commission%20Report%20to%20Congress.pdf [Accessed: Sep. 13, 2023].

[49] A. J. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, et al., *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* 2023. [Online]. Available: arxiv.org/abs/2301.04246. [Accessed: Sep. 13, 2023].

[50] Z. Khanjani, G. Watson, V. P. Janeja, "Audio deepfakes: A survey," *Frontiers in Big Data*, vol. 5, 2023, doi: 10.3389/fdata.2022.1001063.

[51] J. Pu, Z. Sarwar, S. M. Abdullah, A. Rehman, Y. Kim et al., "Deepfake Text Detection: Limitations and Opportunities," *IEEE Symposium on Security and Privacy (SP),* 2023.,doi: 10.1109/sp46215.2023.10179387.

[52] European Parliament's Committee on Culture and Education, *Opinion on the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts*. 2022. [Online]. Available: https://www.europarl.europa.eu/doceo/document/CULT-AD-719637_EN.pdf. [Accessed: Sep. 13, 2023].

[53] C. Rigotti, C. McGlynn, "Towards an EU criminal law on violence against women: The ambitions and limitations of the Commission's proposal to criminalise image-based sexual abuse," *New Journal of European Criminal Law*, vol. 13, no. 4, pp. 452–477, 2022, doi: 10.1177/20322844221140713.

[54] European Commission, *Directive on combating violence against women and domestic violence. Proposal for a Directive of the European Parliament and of the Council on combating violence against women and domestic violence'.* COM(2022) 105 final. 2022.

[55] I. Georgieva, T. Timan, M. Hoekstra, *Regulatory divergences in the draft AI act.* Brussels: Scientific Foresight Unit (STOA), 2022.

[56] European Union, *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).* 2022.

[57] European Commission, *Proposal for a Regulation of the European Parliament and of the Council on the transparency and targeting of political advertising*. COM(2021) 731 final. 2021.

[58] C. Becker, R. Laycock, "Embracing deepfakes and AI-generated images in neuroscience research," *European Journal of Neuroscience*, vol. 58, no. 3, pp. 2657–2661, 2023, doi: 10.1111/ejn.16052.

[59] B. van der Sloot, Y. Wagensveld, B. J. Koops, *Deepfakes: The Legal Challenges of the Synthetic Society,* Tilburg Institute for Law, Technology, and Society, 2021, doi: 10.1016/j.clsr.2022.105716

[60] Europol, *Facing reality? Law enforcement and the challenge of deepfakes, an observatory report from the Europol Innovation Lab*. Brussels: Publications Office of the European Union, 2022.

[61] N. Schick, *Deep Fakes and the Infocalypse*. London: Octopus Books, 2020.

[62] J. Ternovski, J. Kalla, P. M. Aronow, "Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments," *OSF Preprints*, [Online]. Available: https://osf.io/dta97. [Accessed: Sep. 13, 2023].

[63] J. De Cooman, "Humpty Dumpty and High-Risk AI Systems: The Ratione Materiae Dimension of the Proposal for an EU Artificial Intelligence Act," *Market and Competition Law Review*, vol. 6, no. 1, 2022, doi: 10.34632/mclawreview.2022.11304.

[64] L. Edwards, "Regulating AI in Europe: four problems and four solutions", Ada Lovelace Institute, 2022. [Online] Available: https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf. [Accessed: Sep. 13, 2023].

[65] Ada Lovelace Institute, *People, risk and the unique requirements of AI,* 2022. [Online]. Available: https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Policy-briefing-People-risk-and-the-unique-requirements-of-AI-18-recommendations-to-strengthen-the-EU-AI-Act.pdf. [Accessed: Sep. 13, 2023].

[66] Commission Staff Working Document, *Impact Assessment accompanying the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.* 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021sc0084. [Accessed: Sep. 13, 2023].

[67] W. Wahlster, C. Winterhalter, *German Standardization Roadmap on Artificial Intelligence*, DIN, DKE, 2022. [Online]. Available: https://www.din.de/resource/blob/772610/e96c34dd6b12900ea75b460538805349/normungsroadmap-en-data.pdf. [Accessed: Sep. 13, 2023].

[68] N. Helberger, N. Diakopoulos, "ChatGPT and the AI Act," *Internet Policy Review*, vol. 12, no. 1, 2023, doi: 10.14763/2023.1.1682.

[69] T. Mahler, "Between risk management and proportionality: The risk-based approach in the EU's Artificial Intelligence Act Proposal," *Nordic Yearbook of Law and Informatics,* pp. 247–270, 2022, doi: org/10.53292/208f5901.38a67238.

[70] M. C. Sanchez, *Deep fakes: the media and the legal system is under threat*, 2023. [Online]. Available: https://www.lexology.com/library/detail.aspx?g=e4e835cb-f3d1-416e-81b8-81eefe426cf4. [Accessed: Sep. 13, 2023].

[71] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The et al., "Deep Learning for Deepfakes Creation and Detection: A Survey," *SSRN Electronic Journal,* 2022, doi: 10.2139/ssrn.4030341.

[72] V. Cianciaglini, C. Gibson, D. Sancho, O. McCarthy, M. Eira et al., *Malicious Uses and Abuses of Artificial Intelligence*, United Nations Interregional Crime and Justice Research Institute, Europol's European Cybercrime Centre, 2020. [Online]. Available: https://www.europol.europa.eu/cms/sites/default/files/documents/malicious_uses_and_abuses_of_artificial_intelligence_europol.pdf. [Accessed: Sep. 13, 2023].

[73] R. Delfino, "The Deepfake Defense-Exploring the Limits of the Law and Ethical Norms in Protecting Legal Proceedings from Lying Lawyers," Preview Ohio State Law Journal, vol. 84, 2023, doi: 10.2139/ssrn.4355140.

[74] L. Holbrook (2023). *The EU Artificial Intelligence Act and its Human Rights Limitations*. [Online]. Available: https://ohrh.law.ox.ac.uk/the-eu-artificial-intelligence-act-and-its-human-rights-limitations. [Accessed: Sep. 13, 2023].

[75] Centre for Digital Governance, *The false promise of transparent deep fakes: How transparency obligations in the draft AI Act fail to deal with the threat of disinformation and image-based sexual abuse*, Hertie School, 2022. [Online]. Available: https://www.hertie-school.org/en/digital-governance/research/blog/detail/content/the-false-promise-of-transparent-deep-fakes-how-transparency-obligations-in-the-draft-ai-act-fail-to-deal-with-the-threat-of-disinformation-and-image-based-sexual-abuse. [Accessed: Sep. 13, 2023].

[76] K. Nagumotu, "Deep fakes are taking over social media: can the law keep up?," *Intellectual Property Law Review*, vol. 62, no. 2, pp. 102–146, 2022.

[77] M. Ebers, V. R. S. Hoch, F. Rosenkranz, H. Ruschemeier, B. Steinrötter, "The European Commission's Proposal for an Artificial Intelligence Act–A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)," *Multidisciplinary Scientific Journal,* vol. 4, no. 4, pp. 589–603, 2021, doi: 10.3390/j4040043.

[78] R. J. Neuwirth, "Prohibited artificial intelligence practices in the proposed EU artificial intelligence act (AIA)," *Computer Law & Security Review*, vol. 48, 2023, doi: 10.1016/j.clsr.2023.105798.

[79] J. B. Finke, M. F. Larra, M. U. Merz, H. Schächinger, "Startling similarity: Effects of facial self-resemblance and familiarity on the processing of emotional faces," *PLOS ONE*, vol. 12, no. 12, 2017, doi 10.1371/journal.pone.0189028.

[80] T. Dobber, N. Metoui, D. Trilling, N. Helberger, C. de Vreese, "Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?," *The International Journal of Press/Politics*, vol. 26, no. 1, pp. 69–91, 2021, doi: 10.1177/1940161220944364.

[81] T. Nakano, T. Yamamoto, "You trust a face like yours," *Humanities and Social Sciences Communications,* vol. 9, no. 1, 2022, doi: 10.1057/s41599-022-01248-8.

[82] J. Newman, *A Taxonomy of Trustworthiness for Artificial Intelligence*, Center for Long-Term Cybersecurity, Berkeley, 2023. [Online]. Available: https://cltc.berkeley.edu/wp-content/uploads/2023/01/Taxonomy_of_AI_Trustworthiness.pdf. [Accessed: Sep. 13, 2023].

[83] European Parliament's Committee on the Internal Market and Consumer Protection and Committee on Civil Liberties, Justice and Home Affairs, *Draft Report on the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts*. 2022. [Online]. Available: https://www.europarl.europa.eu/doceo/document/CJ40-AM-732836_EN.pdf. [Accessed: Sep. 13, 2023].

[84] E. Morrow, *Beyond disinformation: deep fakes and false memory implantation*, International Neuroethics Society and International Youth Neuroscience Association, Neuroethics Essay Contest, 2021. [Online]. Available: https://www.dana.org/article/neuroethics-essay-general-audience-2021. [Accessed: Jul. 13, 2023].

[85] Bundesrat, *Beschluss des Bundesrates. Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union*. 2021. [Online]. Available: https://www.bundesrat.de/SharedDocs/drucksachen/2021/0401-0500/488-21.pdf?__blob=publicationFile&v=1. [Accessed: Sep. 13, 2023].

[86] European Parliament›s Committee on Legal Affairs, *Report on artificial intelligence: questions of interpretation and application of international law in so far as the EU is affected in the areas of civil and military uses and of state authority outside the scope of criminal justice. 2020/2013(INI)*. 2021. [Online] Available: https://www.europarl.europa.eu/doceo/document/A-9-2021-0001_EN.html. [Accessed: Sep. 13, 2023].

[87] R. Michaelson, *Turkish presidential candidate quits race after release of alleged sex tape*, 2023. [Online]. Available: https://www.theguardian.com/world/2023/may/11/muharrem-ince-turkish-presidential-candidate-withdraws-alleged-sex-tape. [Accessed: Sep. 13, 2023].

[88] L. M. Böswald, B. A. Saab, *What a Pixel Can Tell: Text-to-Image Generation and its Disinformation Potential*, Democracy Reporting International, 2022. [Online]. Available: https://democracyreporting.s3.eu-central-1.amazonaws.com/images/6331fc834bcd1.pdf. [Accessed: Sep. 13, 2023].

[89] P. Beuth, A. Demling, M. Hoppenstedt, T. Kleinz, A. Reiner, et al., "Die große Fake-Maschine," *Der Spiegel*, no. 28, pp. 8–15, 2023.

[90] M. Wong, "We Haven't Seen the Worst of Fake News," *The Atlantic*, 2022. [Online]. Available: https://www.theatlantic.com/technology/archive/2022/12/deepfake-synthetic-media-technology-rise-disinformation/672519. [Accessed: Sep. 13, 2023].

[91] S. Kutzner, "Deepfake: Nein, Jordan B. Peterson zog nicht über Baerbock, Lauterbach und Scholz her," Correctiv, 2023. [Online]. Available: https://correctiv.org/faktencheck/2023/03/14/deepfake-nein-jordan-b-peterson-zog-nicht-ueber-baerbocklauterbach-und-scholz-her. [Accessed: Sep. 13, 2023].

[92] D. Neuerer, T. Stiens, *Wie KI zur Gefahr für die Demokratie werden könnte. Handelsblatt*, 2023. [Online]. Available: https://www.handelsblatt.com/politik/deutschland/deepfakes-wie-ki-zur-gefahr-fuer-die-demokratie-werden-koennte/29221078.html. [Accessed: Sep. 13, 2023].

[93] Center on Extremism, *The Dangers of Manipulated Media and Video: Deepfakes and More*, 2023. [Online]. Available: https://www.adl.org/resources/blog/dangers-manipulated-media-and-video-deepfakes-and-more. [Accessed: Sep. 13, 2023].

[94] Der Standard, *Emma Watson liest „Mein Kampf": Trolle feiern Software für Stimmen-Deepfakes*, 2023. [Online]. Available: https://www.derstandard.de/story/2000143117245/emma-watson-liest-mein-kampf-trolle-feiern-software-fuer-stimmen. [Accessed: Sep. 13, 2023].

[95] J. Habgood-Coote, "Deepfakes and the epistemic apocalypse," *Synthese*, vol. 201, no. 3, 2023, doi 10.1007/s11229-023-04097-3.

[96] D. Leufer, *Access Now's submission to the European Commission's adoption consultation on the Artificial Intelligence Act*, Accessed: Sep. 13, 2023. [Online]. Available: https://www.accessnow.org/wp-content/uploads/2021/08/Submission-to-the-European-Commissions-Consultation-on-the-Artificial-Intelligence-Act.pdf. [Accessed: Sep. 13, 2023].

[97] G. Pennycook, T. D. Cannon, D. G. Rand, "Prior exposure increases perceived accuracy of fake news," *Journal of Experimental Psychology: General*, vol. 147, no. 12, pp. 1865–1880, 2018, doi: 10.1037/xge0000465.

[98] M. Karaboga, "Die Regulierung von Deepfakes auf EU-Ebene: Überblick eines Flickenteppichs und Einordnung des Digital Services Act- und KI-Regulierungsvorschlags," in *Digitale Hate Speech. Interdisziplinäre Perspektiven auf Erkennung, Beschreibung und Regulation*, S. Jaki, S. Steger, Ed. Stuttgart: J. B. Metzler, 2023, doi: 10.1007/978-3-662-65964-9_10..

[99] F. Palmiotto, *Detecting Deep Fake Evidence with Artificial Intelligence A Critical Look from a Criminal Law Perspective*, 2023. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4384122. [Accessed: Sep. 13, 2023].

[100] Washington State Senate, *Washington State Senate Bill on Defining synthetic media in campaigns for elective office, and providing relief for candidates and campaigns.* S.B. 5152. 2023. [Online]. Available: https://lawfilesext.leg.wa.gov/biennium/2023-24/Pdf/Bills/Senate%20Bills/5152-S.E.pdf?q=20230321103533. [Accessed: Sep. 13, 2023].

[101] Center for an Informed Public, "New WA law requires clear disclosures for 'deepfakes' used in election media," [Online]. Available: https://www.cip.uw.edu/2023/06/09/new-wa-law-deepfake-disclosure-election-media. [Accessed: Sep. 13, 2023].

[102] Bitkom, *Täuschend echt, aber alles Lüge: 63 Prozent haben Angst vor Deepfakes*, 2023. [Online]. Available: https://www.bitkom-research.de/news/taeuschend-echt-aber-alles-luege-63-prozent-haben-angst-vor-deepfakes. [Accessed: Sep. 13, 2023].

[103] J. Laux, S. Wachter, B. Mittelstadt, "Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act," *SSRN Electronic Journal*, 2023, doi: 10.2139/ssrn.4365079.

[104] T. Dobber, S. Kruikemeier, F. Votta, N. Helberger, E. P. Goodman, "The effect of traffic light veracity labels on perceptions of political advertising source and message credibility on social media," *Journal of Information Technology & Politics*, 2023, doi: 10.1080/19331681.2023.2224316.

[105] H. Farid, *ChatGPT and Dall-E Should Watermark Their Results*, 2023. [Online]. Available: https://gizmodo.com/chatgpt-dall-e-free-ai-art-should-watermark-results-1850289435. [Accessed: Sep. 13, 2023].

[106] W. Fischer, "California's governor signed new deepfake laws for politics and porn, but experts say they threaten free speech," Business Insider, 2019. [Online]. Available: https://www.businessinsider.com/california-deepfake-laws-politics-porn-free-speech-privacy-experts-2019-10. [Accessed: Sep. 13, 2023].

[107] A. Walorska, *Deepfakes & Disinformation*. Berlin: Friedrich Naumann Fundation, 2020.

[108] J. Langa, "Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes," *Boston University Law Review*, vol. 101, pp. 761–801, 2021. [Online]. Available: https://www.bu.edu/bulawreview/files/2021/04/LANGA.pdf. [Accessed: Sep. 13, 2023].

[109] H. Ajder, J. Glick, *JUST JOKING! Deepfakes, Satire and the Politics of Synthetic Media*, WITNESS, 2021. [Online]. Available: https://cocreationstudio.mit.edu/just-joking/. [Accessed: Sep. 13, 2023].

[110] M. Pawelec, "Deepfakes als Chance für die Demokratie?," in *Digitalisierung und die Zukunft der Demokratie: Beiträge aus der Technikfolgenabschätzung*, A. Bogner, M. Decker, M. Nentwich, C. Scherz, Ed. Baden-Baden: Nomos Verlagsgesellschaft, pp. 89–102, 2022, doi: org/10.5771/9783748928928-89.

[111] F. Gollin, A. Gheorghita, A. Young, O. Ruiz Pilato, X. Chen, *Deepfake. Legal Paper*, Cyber Rights Organization, 2023. [Online]. Available: https://cyberights.org/wp-content/uploads/2023/03/Deepfake-Legal-Paper-CRO2023-1-2.pdf. [Accessed: Sep. 13, 2023].

[112] E. F. Judge, A. M. Korhani, "Deepfakes, Counterfeits, and Personality," *SSRN Electronic Journal,* 2021, doi: 10.2139/ssrn.3893890.

[113] M. Liu, X. Zhang, "Deepfake Technology and Current Legal Status of It," *Proceedings of the 2022 3rd International Conference on Artificial Intelligence and Education (IC-ICAIE 2022)*, 2023, pp. 1308–1314, doi: 10.2991/978-94-6463-040-4_194.

[114] P. Grady, *EU Proposals Will Fail to Curb Nonconsensual Deepfake Porn*, 2023. [Online]. Available: https://datainnovation.org/2023/01/eu-proposals-will-fail-to-curb-nonconsensual-deepfake-porn. [Accessed: Sep. 13, 2023].

[115] N. A. Smuha, "Beyond the Individual: Governing AI's Societal Harm," *Internet Policy Review*, vol. 10, no. 3, 2021, doi: 10.14763/2021.3.1574.

[116] C. Pershan, R. Jonusaite, *User-Guide to the EU Digital Services Act*, EU DisinfoLab, 2022. [Online]. Available: https://www.disinfo.eu/wp-content/uploads/2022/06/20220602_DSAUserGuide_FinalVersion.pdf. [Accessed: Sep. 13, 2023].

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Creating a Repeatable Nontechnical Skills Curriculum for the University of Southern Maine (USM) Cybersecurity Ambassador Program (CAP)

Lori L. Sussman | Department of Technology, University of Southern Maine, USA, ORCID: 0000-0003-3667-0340
Zachary S. Leavitt | Department of Technology, University of Southern Maine, USA, ORCID: 0000-0003-3667-0340

Abstract

The workforce demand for skilled cybersecurity talent has exceeded its supply for years. Historically, the pedagogical approach was to identify and create curricula for the most in-demand technical knowledge, skills, and abilities (KSAs). Unfortunately, the field has tended to neglect nontechnical counterparts. However, recent literature suggests a core set of nontechnical KSAs that employers seek after. This study explored the codification of a nontechnical curriculum for a cybersecurity internship program at the University of Southern Maine (USM). The USM faculty created the Cybersecurity Ambassador Program that can serve students and the community. The service to students is to make them more attractive to employers. The benefit to the community is to provide cybersecurity awareness training to vulnerable populations. This discussion about the USM CAP serves as a case study for other programs considering this type of enrichment using an internship model. CAP started as an informal program, but this research used objective data to create repeatable blueprints. The researchers designed these lesson plans to help students progress from novices to competent in crucial nontechnical skills delineated in the National Initiative for Cybersecurity Education (NICE) Workforce framework. The team used a mixed methods approach to baseline

Received: 23.05.2023

Accepted: 19.09.2023

Published: 19.09.2023

Cite this article as:
L.L. Sussman, Z.S.Leavitt "Creating a Repeatable Nontechnical Skills Curriculum for the University of Southern Maine (USM) Cybersecurity Ambassador Program (CAP)," AICG, vol. 2, no. 1, 2023,
DOI: 10.60097/ACIG/162858

Corresponding author:
Lori L. Sussman,
Department of Technology,
University of South Maine,
USA, ORCID:
0000-0003-3667-0340;
E-MAIL:
lori.sussman@maine.edu

Tier 1/novice students' skill levels, place them in a cybersecurity enrichment program, track their progress, and determine program efficacy in helping them achieve beginner status. The information shared can serve as a point of departure for a case study that might guide other programs interested in doing similar work.

Keywords
NICE Workforce Framework, cybersecurity education, cybersecurity training, cybersecurity ambassador, cybersecurity internships

1. Introduction

IN 2013, like many highly connected nations, the United Kingdom (UK) looked deeply into its cybercrime reports to find actionable trends. Researchers were astonished that network and computer hygiene could prevent 80% of cyberattacks [1, p. 4]. However, the report also found that the workforce needed to be more robust to answer individual and organizational needs. A lack of science and technology courses in UK schools created a workforce gap that would take decades to fill [1, p. 27]. This demand for cybersecurity (CS) talent is not limited to the United Kingdom. In the United States (US), the Cybersecurity and Infrastructure Security Agency (CISA) created the National Initiative for Cybersecurity Careers and Studies (NICCS) to address teacher and student skill shortages. Unfortunately, while these national initiatives helped universities target essential CS knowledge, skill, and abilities (KSAs), they lacked blueprints to create student programs [2]. The hard work of developing repeatable and scalable programs fell on academia to create programs that meet workforce needs.

1.1. Mining for New Talent Pools

The current cybersecurity workforce must be improved to satisfy the demand for qualified cybersecurity professionals. Experts predict this shortfall will continue for several years [3]. As recently as 2018, researchers found that an excess of 1.5 million positions will be unfilled in the global cybersecurity workforce. Businesses seek employees with technical and interdisciplinary credentials to help fill this cybersecurity gap [4]. Given the shortage of qualified cybersecurity professionals, new talent pools of applicants are needed.

Cybersecurity employers historically have overlooked women and people of colour to fill essential roles [5]. In 2021, women comprised more than 50% of the US population, yet, only 35.5% majored in

science, technology, engineering, and mathematics (STEM) disciplines [5]. In addition, the US has a large minority population that is increasing over time but does not enjoy representative numbers in the computing sciences. [6]. Women and minorities can fill this CS workforce gap and should, as they have a vested interest. In 2014, data revealed that a million more US women than men had their identities stolen [7]. On average, people of colour, African American, and people of Latino descent are two to three times more likely than white people to become victims of fraud related to debt or income [8].

Moreover, existing security technologies disadvantage women and people of colour. For example, biometric facial recognition systems have trouble identifying the faces of women and people of colour [9]. Therefore, increasing women and minorities in cybersecurity enlarges the talent pool and provides new perspectives to improve technologies and practices within the field.

1.2. Next Generation Cyber Professional Curriculum Development

Hiring managers created job descriptions that screen potential employees for cybersecurity skills in various tools and systems. As such, candidates who made it to the interview stage often had comparable technical skill sets. However, it was often nontechnical, called soft skills, that got the candidate the job [10]. Industry, government, and academia members noted that CS graduates frequently lacked the necessary soft, hard, and mixed nontechnical KSAs for employment [11]. Employers almost universally recognize that entry-level workers need client-facing KSAs to accomplish the organization's cybersecurity goals. KSAs involving written and oral communication, teamwork, problem-solving, and critical thinking skills were particularly important [10, 11].

The impetus for the Cybersecurity Ambassador Program (CAP) program started with University of Southern Maine (USM) cybersecurity students who wanted to do community outreach. The faculty worked with the State of Maine Office of Securities and the Maine Economic Initiative Fund (MEIF) to secure initial grant funding to make this community engagement program a paid departmental internship. As part of the grant requirements, the sponsor asked the faculty to create a program that served students and the community. There was also a requirement to research the efficacy of the approach. After securing approval from the USM Institutional Review Board (IRB), the faculty moved forward to create a model that could serve as a case study for other cybersecurity educational organizations.

The faculty recruited students and prioritized those who had Federal Work Study (FWS) funding and students who needed to take the mandatory internship class for their program of study. FWS students usually worked ten hours per week, and interns worked twenty hours per week during a sixteen-week semester. The 2019 cohort started with two students but was cut short due to COVID restrictions. The primary advisor had to recast the program as entirely virtual for the next three years. The program involved fourteen students in seven cohorts (Tab. 1).

Table 1. Student Participants.

Undergraduate Students in Cohorts	
Fall 2019	2
Fall 2020	2
Spring 2021	4
Fall 2021	5
Spring 2022	5
Fall 2022	7
Spring 2023	7
Graduate Assistants	
AY20-21	1
AY21-22	3
AY22-23	3

The emphasis on community service by providing cybersecurity awareness training to vulnerable populations is an ideal vehicle to focus on needed nontechnical KSA development. CAP promoted CS awareness and education through research and outreach opportunities that, in turn, required students to elevate communications and leadership skills. The program leveraged undergraduate and graduate students seeking to make meaningful contributions to local communities as the students gained vital professional competencies.

1.3. Skill Acquisition

Proficiency via objective assessment was critical. The program used stages from the Five-Stage Model of Adult Skill Acquisition as its conceptual model [12]. This framework describes how people learn skills and identifies five stages of progress novice, advanced beginner, competent, proficient, and expert. The faculty presumes students enter the program at the novice level, defined as the Bronze/Tier 1 stage. The blueprint for the program scaffolded instruction sequentially and progressively while allowing participant autonomy to consume the content. As Dreyfus noted, "The student needs not only the facts but also an understanding of the context in which that information makes sense" [12, p. 177]. The intention was to construct a professional development journey for these students to progress to advanced beginners, corresponding to the Silver/Tier 2 proficiency, a second sixteen-week program. This advanced beginner signpost stage, characterized by exposure to sufficient examples of meaningful activities, is critical for students' ability to apply learning to new and novel situations. When students achieve mastery of the silver curriculum, they have reached the competent stage, commemorated by promotion to Gold/Tier 3 status.

2. Methods

The researchers used a phenomenological study approach with structured and semi-structured data collection methods. Over sixteen weeks for this pilot, the researchers observed students, assessed their assignments, conducted interviews, and objectively evaluated participants performing various internship activities. The objective was to provide students with enhanced education, experience, and exposure to cybersecurity awareness and training research. The researchers used US Commerce's National Institute for Standards and Technology (NIST) National Initiative for Cybersecurity Education (NICE) – specific nontechnical KSAs in a focused way for curriculum development. The intention was to enhance 19 nontechnical skills from the NICE Workforce Framework using student-led projects for community awareness training. External assessment of the program was uniformly positive. For example, the National CyberWatch Centre identified this program as the 2021 Cybersecurity Curriculum Best Innovation. The faculty also received several other awards from the Epsilon Pi Tau (EPT), a technology honour society. The primary investigator won the EPT Warner award in 2021, 2022, and 2023 for presentations to the Cybersecurity Ambassadors. Student participants also received numerous internal USM recognition and received job offers at faster rates than nonparticipating students. As such, this program design can serve as a valuable case

study for other academic institutions interested in similar student enrichment opportunities.

2.1. Program Design

The program used standard job site internships to provide opportunities for concurrent college credit. CAP created a cohort of student interns each academic semester who had demonstrated interest/ability in cybersecurity career pathways. The Principal Investigator (PI) created a three-tier program. The Cybersecurity Ambassador (CA) students achieved the first or bronze tier through oral and written assessments based on technical training and career planning modules developed and taught by working cybersecurity researchers and cybersecurity professionals. The second (silver) phase emphasized leadership and mentorship. The PI and Coordinator assessed student outreach leadership and peer training/mentorship, incorporating data into cybersecurity awareness and training research efforts. The top (gold) level was where students functioned at the programmatic level, helped coordinate outreach, and certified other students. These tiers correspond to novice, beginner, and advanced beginner levels of expertise.

The program gave participants paid entry-level cybersecurity internships, which allowed them to use the class for their program's mandatory internship requirement. The CAP created a cohort of student interns each academic semester who founds ways to learn nontechnical skill mastery as they pursued Cybersecurity career pathway material. The Principal Investigator (PR) intended to create a three-tier program to incentive continued participation and skill mastery. The Cybersecurity Ambassador (CA) students achieved the first or bronze tier through oral and written assessments based on technical training and career planning modules developed and taught by working Cybersecurity researchers and Cybersecurity professionals. The second (silver) phase emphasized leadership and mentorship. The PI and Coordinator assessed student outreach leadership, peer training/mentorship, incorporating data into cybersecurity awareness and training, and their research efforts. The top (gold) level was where students functioned at the programmatic level, helped coordinate outreach, and certified other students.

2.2. Nontechnical Skill Curriculum Development Process

The faculty advisor that founded CAP derived requisite nontechnical KSAs for cybersecurity students using those listed in the National Initiative for Cybersecurity Education (NICE) Workforce

Framework. Graduate Assistants used these KSAs to formulate learning objectives and activities for bronze-level students to access in their weekly skill enrichment modules in the university learning management system [10] (Tab. 2).

Table 2. These are the 19 NICE nontechnical competencies from the Workforce Framework. Graduate students and faculty create activities and assess mastery. Note students have three signpost stages. KSAs were scaffolded to be sequential and progressive.

KSA's	Curriculum Map
Presentation skills	Bronze (Tier 1)
Developing positive customer relations	Bronze (Tier 1)
Written communications skills	Bronze (Tier 1)
Working effectively with peers	Bronze (Tier 1)
Intellectual curiosity	Bronze (Tier 1)
Using computers effectively	Bronze (Tier 1)
Adaptability	Bronze (Tier 1)
Professional demeanour	Bronze (Tier 1)
Training	Bronze (Tier 1)
Ethics in decision making	Bronze (Tier 1)
Managing personal stress	Bronze (Tier 1)
Customer Service Problem Resolution	Silver (Tier 2)
Knowledge of core business processes	Silver (Tier 2)
Knowledge of and compliance with legal and regulatory requirements	Silver (Tier 2)
Managing crises	Silver (Tier 2)
Critically using information for decision making	Gold (Tier 3)
Facilitating teams and teamwork	Gold (Tier 3)
Negotiating techniques	Gold (Tier 3)
Leadership abilities	Gold (Tier 3)

Initially, observation was the predominant evaluation mechanism to assess student mastery. However, the need to codify the CAP curricula became apparent as the program grew. The team developed research approaches that supported the university's cybersecurity credentialing, provided a repeatable curriculum that students enjoyed, and met the goals of enriching skills that made participants particularly attractive workforce candidates. This vision spawned research questions that focused on nontechnical skill attainment for students with some STEM and cybersecurity background and included in the Bronze/Tier 1 curriculum:

1. How can the CAP objectively assess student baseline KSA of novice (Bronze/Tier 1) KSAs?

2. How can the team craft maximum flexibility into the Bronze/Tier 1 curriculum to accommodate different paces for participants?

3. How can the program objectively assess nontechnical KSA mastery progress from novice (Tier 1) to beginner (Tier 2)?

This rigor provided measurable data on students' progress and improved learning techniques.

The team created a 16-week curriculum designed to increase the speed of acquisition and retention of nontechnical KSAs by bronze-level ambassadors enrolled in the CAP. This research is a snapshot of the pilot group. The team received and filled out a baseline and a weekly survey to measure ambassadors' perceptions of their KSA development. The team also asked participants to suggest improvements to curriculum areas.

In this survey, the team asked the ambassadors to respond to each of the following three statements using a Likert scale from "strongly disagree" to "strongly agree":

- First, the content was relevant to the weekly learning objectives.
- The content was well-organized and easy to understand.
- Finally, the learning activities and assessments were effective in reinforcing the content.

Additionally, the team asked ambassadors to identify the best aspect and most challenging parts of each module. The survey incorporated a text area to capture feedback for these two questions and the question, "What is one thing we could do better for the next group?"

The feedback provided by the bronze-level ambassadors provides opportunities for future refinement and improvement of the curricula.

This qualitative data collection used a pre- and post-curriculum self-evaluation survey to measure ambassadors' perceptions. The objective was to capture the bronze-level ambassadors' journey from novice to advanced beginner. The survey instrument used a Likert scale to measure the ambassadors' confidence in their abilities associated with the targeted KSA.

The researchers also quantitatively measured proficiency via quizzes, discussion posts, and assignments graded by rubrics. These assessment techniques allowed for objectively validating mastery for KSAs at the novice level. Table 3 shows the mix of objective tools used.

Table 3. Objective Measures of KSA Proficiency.

Quiz	Handbook Knowledge
Quiz	Professional Demeanour
Quiz	Ethics in Cybersecurity
Graded Assignment	Article Review
Graded Assignment	Handout Creation
Graded Assignment	Presentation Deck Creation
Graded Assignment	Progress Reports
Graded Assignment	Discussion Posts
Graded Assignment	After Event Reflection

The researchers began the instruction process by decomposing the task environment into context-free features that the beginner could accomplish without the desired skill [12]. The learning management system (LMS) had posted rules for the Bronze Ambassadors, which allowed them to navigate the curriculum through self-paced modules. The students used their internship time to consume materials, complete assignments, and get feedback to improve results.

2.3. Assumptions, Limitations, and Scope

Several significant limitations impacted the design and outcomes of this study. The two biggest challenges were time and money. Academic semesters are typically sixteen weeks, but

administrative items such as onboarding, vacation, and finals consume one to three weeks. Therefore, the designers had to condense the program to 13 – 14 weeks. Also, not all Ambassadors had the same number of hours each week. Because funding for CA pay came from both internship grants and the Federal Work-Study Program, half the Ambassadors worked ten and about half worked 20 hours per week. As such, the program had to have a flexible and achievable design to meet the program-related objectives.

The team balanced new content creation by incorporating pre-existing content for each curriculum module. Content specifically tailored for CAP use was ideal. However, creating content for each module described in the following section proved too expensive, time-consuming, or both. As such, the 16-week CAP curricula included newly built and previously published content. When possible, the team used free online e-learning platforms and tools like LinkedIn Learning or Coursera since they had built-in ways for students to prove their proficiency.

The team delivered the program entirely online, using Brightspace. Use of an LMS assisted in monitoring and assessing the effectiveness of the CAP curricula's objectives. The weekly curriculum survey also offered an opportunity to solicit any access issues. Additionally, CAs could raise concerns asynchronously via the CAP Discord tool.

The asynchronous delivery allowed students access to the course materials at their convenience. Supplemental material was available to CAs via Brightspace and a shared CAP Google Drive. The team also used external content from platforms such as YouTube, but the downside was that this approach relied on provider availability of the provider for access. The team created as much video content as feasible during this sixteen-week session.

The research team recorded all Brightspace data, including student progress and assessments, in a separate Google Drive accessible only to the research team. In addition to this data, the researcher also collected qualitative data from Qualtrics survey results. This survey captured participant self-evaluations. This multi-faceted approach to data collection allowed the team to understand the ambassadors' experiences with the curricula and evaluate the efficacy of the nontechnical KSAs developed through the program.

2.4. Participants/Sample

The team employed purposive sampling to select participants who met specific criteria. The initial sample consisted of four

bronze-level CAs hired onto the spring 2023 CAP cohort. However, one participant withdrew from the program before its completion, leaving three participants for the study. The participants were undergraduate students who had not yet gained significant work experience in the CS field. They were selected based on their potential to develop the nontechnical KSAs identified in the study. Additionally, the participants represented historically underrepresented demographics in the CS and IT disciplines.

The three participants were diverse regarding their demographic backgrounds, including gender and ethnicity. All participants had completed introductory courses in their cybersecurity program and expressed interest in pursuing a career. The participants were highly motivated and committed to developing their skills in the CAP program.

Pilot participant demographics were relatively diverse. Of the ambassadors who completed the pre-survey, 75 percent identified as Caucasian, and the remaining 25 percent identified as Black. Data showed an even split between male and female participants. About half of the participants had earned a bachelor's or associate's degree, while the other half had some college education but no degree. It is worth noting that none of the cohorts had previous experience in the field of CS or had served in the US Armed Forces. As a result, the participants in this study represent a historically underrepresented group in the CS workforce and highlight the need to diversify the field.

2.5. Data Collection

The CAP used free software or software provided at no cost to students enrolled in the UMS. This software included programs from Brightspace Learning Management System (LMS), Microsoft 365 (Word, Excel, PowerPoint, etc.), Google Workspace (Gmail, Slides, Drive, etc.), and standalone programs like Canva, Discord, Trello, Zoom, and Zotero. The research team required the CAs to agree to the end-user license agreements (EULAs) and privacy policies of the software mentioned above. Researchers did not gather data from tool use. Instead, qualitative exploration using the web application Brightspace and a Qualtrics survey were collection tools.

Brightspace is USM's LMS and offers features to conduct online assessments, host discussion forums, and deliver remote instruction to students. Students were familiar with the tool before joining CAP–the study design generated data from participants' interactions with the content disseminated via Brightspace. Like the above software, students had to accept the EULA and data privacy policy for Brightspace

before enrolling in USM's online courses. By extension, participants in the 16-week CAP curricula had to opt into the Brightspace privacy policy to participate in the study.

Instead of using personally identifiable information (PII) such as the ambassadors' first or last name or email address, the team coded and anonymized all student data. Researchers used randomly generated identification numbers for all participants. As with Brightspace, students who used Google Drive, Gmail, or Slides had to accept the privacy policies and any EULAs of that software before participating in the study.

3. Results

3.1 Qualitative Results

During the study, researchers discovered that not all participants provided feedback every week due to the modules' omission in weeks 10 through 15. This erratic feedback was a Brightspace survey limitation because the system mapped them in advance and did not update with an evolving curriculum. Additionally, the initial implementation of the survey radio option in Brightspace improperly grouped data. That idiosyncrasy affected the count of each value on any question using a Likert scale.

For instance, one of the survey questions asked participants to rate their agreement with the statement "The content was relevant to the weekly learning objectives" on a scale of 1 to 5, with one being strongly disagreed and five being strongly agreed. While the survey accurately captured participants' written responses, the data grouping was inaccurate, which could have impacted the analysis of the data (Fig. 1).

#	Statement	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	The content was relevant to the weekly learning objectives.	◯	◯	◯	◯	◯
2	The content was well-organized and easy to understand.	◯	◯	◯	◯	◯
3	The learning activities and assessments were effective in reinforcing the content.	◯	◯	◯	◯	◯

Figure 1. Likert Scale Questions in the Brightspace Survey. It demonstrates a partial evaluation of the Brightspace curriculum modules. This figure is an April 1st, 2023 snapshot.

To address the data grouping issue, the team created individual surveys for each curriculum module and released them to participants who had not yet provided feedback for weeks 10 through 15. This approach allowed for the accurate capture of data related to each module. The team also created individual surveys for weeks one through eight but hid them from users to prevent data duplication and use in future semesters. Individual surveys for each module allowed for a more detailed analysis of the feedback received, which will inform the ongoing development of the CAP curriculum.

The purpose of this study was to evaluate the impact of a cybersecurity professional development curriculum on the skills and knowledge of participants. The pre-and post-curricula measured changes in confidence levels and understanding of various KSAs. The results of the pre-curricula survey indicate that participants generally had a moderate level of confidence in their presentation, written communication, negotiation, and crisis management abilities, with mean scores ranging from 3.25 to 4.5 out of 5. Participants also demonstrated a moderate understanding of cybersecurity compliance and legal and regulatory requirements, with a mean score of 3.25 out of 5. However, participants reported lower confidence levels in their ability to resolve cybersecurity problems and seek out cybersecurity news and information, with mean scores of 2.75 and 3 out of 5, respectively.

After completing the cybersecurity professional development curriculum, participants reported significant improvements in their confidence levels and understanding of various skills and knowledge domains (Fig. 2).

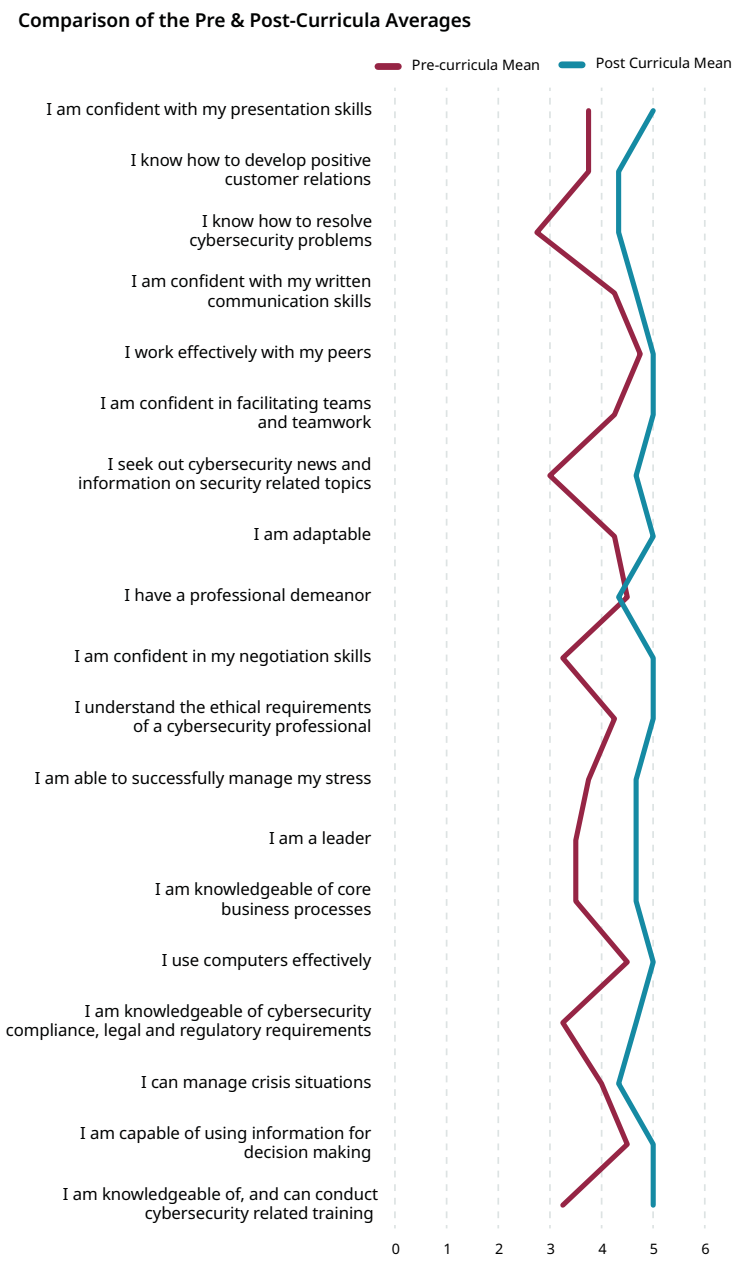

Figure 2. Comparison of the Pre- and Post-Curricula Averages. It depicts a comparison of the change of averages between the pre-and post-curricula KSA inventory surveys.

The post-curricula survey results show that participants' confidence levels increased significantly in all domains, with mean scores ranging from 4.33 to 5 out of 5. Participants also reported a substantially

higher understanding of cybersecurity compliance and legal and regulatory requirements, with a mean score of 4.67 out of 5. Moreover, participants reported a significant improvement in their ability to resolve cybersecurity problems, seek out cybersecurity news and information, and effectively work with peers, with mean scores ranging from 4.33 to 5 out of 5 (Fig. 2).

3.2. Quantitative Results

The quiz design included a pool of questions and displayed ten randomly to the student. There was no time limit, but they only got one attempt. Assignments had a rubric to assure objective and consistent grading by faculty and graduate assistants, but the feedback was either "pass" or "redo" (Tab. 4).

Table 4. Pilot Results.

Assessment	Title	Student A	Student B	Student C
Quiz	Handbook Knowledge	97.23%	100.00%	97.23%
Quiz	Professional Demeanour	100.00%	80.00%	90.00%
Quiz	Ethics in Cybersecurity	60.00%	50.00%	50.00%
Graded Assignment	Article Review	Pass	Pass	Pass
Graded Assignment	Handout Creation	Pass	Pass	Pass
Graded Assignment	Presentation Deck Creation	Pass	Pass	Pass
Graded Assignment	Progress Reports	Pass	Pass	Pass
Graded Assignment	Discussion Posts	Pass	Pass	Pass
Graded Assignment	After Event Reflection	Pass	Pass	Pass

The students showed consistency of mastery of the handbook and professional demeanour content. However, there was consistent underperformance in the cybersecurity ethics topic.

4. Discussion

This study assessed the CAP curricula' effectiveness in developing targeted nontechnical KSAs among bronze-level ambassadors using qualitative and quantitatively collecting instruments. The team formulated three research questions to guide the study toward its intended purpose. First, the results indicated that the 16-week bronze-level CAP curricula enhanced participants' skills and knowledge in various domains. In addition, the significant improvements in their confidence levels and understanding of the topics indicated the move from novice to advanced beginner stages. Based on these findings, the data suggests that this program should continue to use this repeatable process and start working on the curriculum to help students move from beginner to advanced beginner signpost stages.

4.1. Qualitative Findings

The qualitative data indicated that the participants saw the opportunity to work on meaningful projects, have clear objectives for evaluation, and get practical experience as the most valuable aspect of the internship. The feedback from their self-assessments indicated increased confidence in all graded areas. The slight declination in professional demeanour may have been due to the combination of low objective quiz scores despite high pass rates on the assignments. Regardless, these students uniformly appreciated the chance to apply what they learned in the classroom to real-world situations and work alongside experienced industry professionals. To this point, one student said the following,

> *This experience has contributed to enhancing my professional attitude, and as a result, my self-perception has changed. I now have increased confidence in presenting in public and explaining cybersecurity terms to nontechnical individuals, which means I have gained confidence in my ability to communicate effectively with diverse audiences.*

The survey feedback indicated that this internship program provided an excellent work culture with sufficient supervision and feedback to grow. The qualitative data showed that CAP provided a supportive and inclusive work environment, and the curriculum helped gain future employment.

4.2. Quantitative Findings

There could be many reasons the students performed well in the first two quiz areas but not the third. It could be due to differences in interest, motivation, learning style, or prior knowledge. For example,

Student A might have a strong interest in Handbook Knowledge and Professional Demeanour content and might have prior knowledge or experience in these areas. On the other hand, Student A might have less interest or prior knowledge of cybersecurity ethics.

The assessment may not be a good measure of student learning or mastery. For example, the evaluation might not align with the learning objectives or might not be measuring the right skills or knowledge. For this reason, the research team is reworking the ethics module and quiz questions.

Finally, the underperformance could explain why students self-reported less confidence in their professional demeanour at the end of the program. The researchers did categorize ethics as a subcategory for the students. They may have seen their lack of quantitative performance on the quiz as a reason to question their mastery of content in this area. The interviews indicate the plausibility of this explanation. The student feedback was that the experiences on the quiz and with different vulnerable populations made them self-aware that they had more to learn. The researchers discovered that they must make professional demeanour mastery at the novice, beginner, and advanced beginner levels clearer to students. In this case, they had novice mastery. Still, they expected a higher level of skills usually commensurate with Silver/Tier 2 level, thus reporting a decrease in anticipation of further skill mastery in this area.

4.3. Findings Based on Mixed Methods

The first research question explored the creation of an objective baseline instrument for participants. The one developed using the NICE Workforce Framework, and the Five-Stage Model of Adult Skill Acquisition provided an effective tool. The data from this survey showed that students self-identified as having a solid foundation in various CS KSAs, according to the comparison of the pre-and post-curricula surveys of bronze-level ambassadors enrolled in corresponding CAP curricula. The pre-curricula poll revealed that most students already felt confident and proficient in these subjects, with mean values above 3.5 for most skills. However, the post-curricula survey showed that the students' confidence and proficiency in some of these areas had increased even further, with mean values for several skills rising above 4.5.

The higher variability in some skills, such as leadership and core business processes, suggests that some students needed more practice in these areas. In addition, while the post-curricula survey showed

an improvement in confidence and proficiency for these KSAs, their mean values remained below 4.5, indicating that there is still room for improvement.

The second question dealt with curriculum flexibility. The conceptual model allowed the team to develop questions to identify critical nontechnical KSAs that help ambassadors progress from novice to beginner and evolve the content accordingly. The data analysis shows that the CAP curricula design, which included written communication, customer service, and stress management, helped develop these critical skills despite students being at different skill levels and needing to absorb the content at different speeds. Students developed skills through lectures, reading, and hands-on activities. This experiential learning provided students with a comprehensive understanding of concepts and necessary nontechnical KSAs when they were most available to absorb the content.

The third research question explored objective measuring approaches for monitoring students' progress from novice to advanced beginner. The findings show that the 16-week CAP curricula used various assessment methods to measure student progress, including formative assessments, quizzes, and activities. These assessments helped to evaluate students' mastery of the nontechnical KSAs covered in the curriculum.

Overall, the data analysis shows that the CAP was influential in developing nontechnical KSAs among students. The program's curriculum design, assessment methods, and andragogical learning elements helped foster these skills development. Therefore, the researchers achieved the purpose of the study, which was to assess the effectiveness of the CAP curricula in developing targeted nontechnical KSAs among bronze-level ambassadors.

However, it is vital to acknowledge the study's limitations and discrepancies in its findings. For example, the study's small sample size limits generalizability. Additionally, the single geographical location may also be a limiting factor. Nonetheless, the findings provide valuable insights into the program's effectiveness in developing nontechnical KSAs and can inform the development of similar programs in the future.

5. Conclusions

The study's findings have several implications for individuals and organizations involved in CS education, training, and awareness.

First, the results provide insight into the effectiveness of the 16-week bronze-level CAP curriculum in developing nontechnical KSAs among bronze-level ambassadors and emphasize the importance of addressing the current CS skills gap.

The data analysis showed that the CAP curriculum design, assessment methods, and elements of andragogical learning could help bridge the CS skills gap and equip individuals with necessary nontechnical KSAs. This finding is significant for individuals seeking to improve their nontechnical KSAs and organizations and institutions seeking to develop the next generation of the CS workforce.

In addition, the study's findings have implications for transformative learning and leading. For example, the CAP curriculum design, which focused on communication, customer service, and stress management, helped foster the development of students' critical thinking and problem-solving skills. These skills are essential for transformative learning and leading, enabling individuals to address complex problems and make informed decisions.

The data generated showed that the demand for CS skills continues to grow. The Bureau of Labour Statistics projects that "employment in computer and information technology occupations is projected to grow 15 percent from 2021 to 2031, much faster than the average for all occupations" [13]. This data highlights the need for individuals to develop and hone nontechnical KSAs to remain competitive in the job market.

Furthermore, the study's findings contribute to the larger literature, knowledge, and practice in CS education and training. Finally, the results provide insights into the effectiveness of nontechnical KSAs in addressing the CS skills gap and offer recommendations for developing similar programs in the future.

This study's findings have practical implications for individuals, communities, organizations, and institutions involved in CS education and training. The results highlight the importance of developing and honing nontechnical KSAs and offer insights into practical methods. The data generated from the proposal also underscore the need for individuals to acquire these skills to meet the growing demand for CS talent. Additionally, the study's findings have implications for transformative learning and leading and contribute to the larger body of literature, knowledge, and practice for CS education and training. For this reason, CAP provides an excellent point of departure for other academic institutions interested in starting their version of the program.

5.1. Implications

The study's findings have several implications for individuals and organizations involved in CS education, training, and awareness. First, the results provide insight into the effectiveness of the 16-week Bronze-level/Tier 1 CAP curriculum in developing nontechnical KSAs among novice students by addressing the current CS skills gap. The team used skills verified by compliance organizations and hiring managers.

Data analysis showed that the CAP curriculum design, assessment methods, and elements of andragogical learning could help bridge the CS skills gap and equip individuals with necessary nontechnical KSAs. This finding is significant for individuals seeking to improve their nontechnical KSAs and organizations and institutions seeking to develop the next generation of the CS workforce. The team also noted a unifying effect that served to nurture students from underrepresented populations – that sense of purpose and belonging supported undergraduate cybersecurity classes where they were the minority.

In addition, the study's findings have implications for transformative learning and leading. For example, the CAP curriculum design, which focused on communication, customer service, and stress management, helped foster the development of students' critical thinking and problem-solving skills. These skills are essential for transformative learning and leading, enabling individuals to address complex problems and make informed decisions. As a result, CAP students achieved student leadership recognition at disproportionately higher rates than their peers.

The data generated from the proposal also showed that the demand for CS skills continues to grow. For example, the Bureau of Labour Statistics projects that "employment in computer and information technology occupations is projected to grow 15 percent from 2021 to 2031, much faster than the average for all occupations" [13]. This employment data highlights the need for individuals to develop and hone nontechnical KSAs to remain competitive in the job market.

Furthermore, the study's findings contribute to the larger body of literature, knowledge, and practice in CS education and training. Finally, the results provide insights into the effectiveness of nontechnical KSAs in addressing the CS skills gap and offer recommendations for developing similar programs in the future.

This research has practical implications for individuals, communities, organizations, and institutions involved in CS education and

training. The results highlight the importance of developing and honing nontechnical KSAs and offer insights into practical methods. The data generated from the proposal also underscore the need for individuals to acquire these skills to meet the growing demand for CS talent. Additionally, the study's findings have implications for transformative learning and leading and contribute to the larger body of literature, knowledge, and practice in CS education and training.

5.2. Recommendations

Based on the conclusions drawn from the findings of this study, the following recommendations may improve CS education and training. First, more organizations must develop and implement similar CS education and training programs. Second, organizations and institutions involved in CS education and training should develop and implement programs targeting nontechnical KSAs, such as communication, customer service, stress management, and critical thinking. The 16-week CAP curriculum design, assessment methods, and elements of andragogical learning can serve as a model for developing and implementing such programs aimed at developing these nontechnical KSAs in students with varying degrees of experience. Third, researchers should conduct more studies to adapt such programs based on workforce demands.

Similarly, one cannot overstate the importance of including nontechnical KSAs in CS education and training. CS education and training programs should include nontechnical KSAs to enhance the effectiveness of their training programs. Incentives such as micro-learning and micro-credential opportunities are motivational and ways to assure assessment credibility. Organizations such as the Educational Design Lab (EDL) offer targeted micro-badges, including collaboration, creative problem-solving, critical thinking, oral communication, and resilience, which helped form the CAP curricula assessment tools.

Finally, the program fosters unique collaboration and peer learning, keeping students on track to graduate on time. Organizations and institutions could foster similar cooperation and camaraderie among students. A CAP-like program promotes peer learning, enhancing the effectiveness of its CS education and training programs. Peer learning opportunities, such as those provided by the CAP Podcast and the CAP Cyber Bowl, can promote leadership opportunities and collaborative engagement while developing students' nontechnical KSAs.

5.3. Reflections

This study explored developing and implementing a sixteen-week CAP curriculum targeting nontechnical KSAs. The data showed that this program might effectively improve the acquisition of nontechnical KSAs in participating students. The study highlights the importance of nontechnical KSAs in building a culture of CS and bridging the CS skills gap. The findings suggest that the CAP curriculum can effectively develop students' nontechnical KSAs related to CS and contribute to improved CS awareness and behaviour, ultimately protecting organizations, customers, and assets from cyber threats. This enrichment for any student, especially those from underrepresented populations, provides the cybersecurity workforce with an entry-level worker performing at a higher-than-expected level of competency.

The baseline and progress made by the pilot group showed that nontechnical KSAs play a critical role in cybersecurity education and student confidence. This study underscores the importance of nontechnical KSAs in cybersecurity education and highlights the potential of education and training to bridge the cybersecurity skills gap. In addition, the findings suggest that developing nontechnical KSAs can improve CS awareness and behaviour, ultimately protecting organizations, customers, and assets from cyber threats.

The exploration of curriculum flexibility was a critical aspect of program success. Developers effectively developed content that provided students with a CS context for these nontechnical KSAs. In addition, the sequential and progressive nature of the modules allowed students to apply learning in novel situations. For example, they could improvise while presenting, thus leading to positive customer relations, better use of computers and computing tools, fluid presentation skills, better-written communication skills, working more effectively with peers, adaptability, intellectual curiosity, managing personal stress, and maintaining a professional demeanour. These skills are essential for individuals seeking to pursue a career in cybersecurity and line up well with the next tier as Silver/Tier 2 Cybersecurity Ambassador.

Finally, the CAP curriculum can be adapted and applied in different learning environments. The study highlights the potential of the CAP curriculum to be adapted and used in K-12, community colleges, or technical schools. Additionally, a micro-credential would add incentives and benefit individuals and employers by providing a clear and recognized standard for evaluating job candidates' nontechnical CS skills.

As the literature indicated, the workforce demand for skilled cybersecurity talent has exceeded its supply for numerous consecutive years. Historically, the pedagogical approach was to identify and create curricula for the most in-demand technical knowledge, skills, and abilities (KSAs). However, recent research suggests adding a core set of nontechnical KSAs that employers seek after. This study explores the codification of a nontechnical curriculum for a cybersecurity internship program at the University of Southern Maine (USM). The USM faculty created the Cybersecurity Ambassador Program to serve students and the community. The service to students was to make them more attractive to employers. The benefit to the community was to provide cybersecurity awareness training to vulnerable populations. This discussion about the USM CAP serves as a case study for other programs considering this type of enrichment using an internship model.

CAP started as an informal program but needed repeatable blueprints. The researchers designed these lesson plans to help students progress from novices to competent in crucial nontechnical skills delineated in the National Initiative for Cybersecurity Education (NICE) Workforce framework. The team used a mixed methods approach to baseline Tier 1/novice students' skill levels, place them in a cybersecurity enrichment program, track their progress, and determine program efficacy in helping them achieve beginner status. The information shared can serve as a point of departure for a case study that might guide other programs interested in doing similar work. Overall, this study offers valuable insights into the effectiveness of the CAP curriculum and suggests promising areas for further research and development in cybersecurity education and training. In addition, it provides a valuable contribution to cybersecurity education and training, with potential benefits for individuals, organizations, and communities alike.

Funding

The Maine State Office of Securities, the Maine Technologies Institute, and the Maine Economic Initiative Fund supported this work.

References

[1] V. Marshall, L. Mills, J. Weingard, J. Young, The UK cyber-security strategy: Landscape review, National Audit Office, United Kingdom, 2013. [Online]. Available: https://www.nao.org.uk/reports/the-uk-cyber-security-strategy-landscape-review/. [Accessed: Sep. 19, 2022].

[2] M.E. Armstrong, K.S. Jones, A.S. Namin, D.C. Newton, "Knowledge, skills, and abilities for specialized curricula in cyber defense: Results from interviews with cyber professionals," *Association for Computing Machinery (ACM) Transactions on Computing Education*, vol. 20, no. 4, pp. 1 – 25, 2020, doi: 10.1145/3421254.

[3] K. Cabaj, D. Domingos, Z. Kotulski, A. Respício, "Cybersecurity education: Evolution of the discipline and analysis of master programs," *Computers & Security*, vol. *75*, pp. 24 – 35, 2018, doi: 10.1016/j.cose.2018.01.015.

[4] J. Peeler, "(ISC)² Study: Workforce Shortfall Due to Hiring Difficulties Despite Rising Salaries, Increased Budgets and High Job Satisfaction Rate," (ISC)², 2015. [Online]. Available: https://blog.isc2.org/isc2_blog/2015/04/isc-study-workforce-shortfall-due-to-hiring-difficulties-despite-rising-salaries-increased-budgets-a.html. [Accessed: Nov. 1, 2022].

[5] D.N. Burrell, "An exploration of the cybersecurity workforce shortage," in *Cyber Warfare and Terrorism: Concepts, Methodologies, Tools, and Applications*, Management Association, Information Resources Ed. Hershey, PA: IGI Global, 2020, pp. 1072 – 1081, doi: 10.4018/978-1-7998-2466-4.

[6] R.T. Palmer, R.J. Davis, T. Thompson, "Theory meets practice: HBCU initiatives that promote academics success among African Americans in STEM," *Journal of College Student Development*, vol. 51, no. 4, pp. 440 – 443, 2010, doi: 10.1353/csd.0.0146.

[7] WR Poster, "Cybersecurity needs women," *Nature*, vol. *555*, no. 7698, pp. 577 – 580, 2018, doi: 10.1038/d41586-018-03327-w.

[8] Federal Trade Commission Report to Congress, *Combating Fraud in African American and Latino Communities*, 2016. [Online]. Available: https://www.ftc.gov/system/files/documents/reports/combating-fraud-african-american-latino-communities-ftcs-comprehensive-strategic-plan-federal-trade/160615fraudreport.pdf. [Accessed: May 22, 2023].

[9] C.M. Cook, J.J. Howard, Y.B. Sirotin, J.L.Tipton, A.R. Vemury, "Demographic Effects in Facial Recognition and Their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, no. 1, pp. 32 – 41, 2019, doi: 10.1109/TBIOM.2019.2897801.

[10] L.L. Sussman, "Exploring Nontechnical Knowledge, Skills, and Abilities (KSA) that May Expand the Expectations of the Cyber Workforce," *Cybersecurity Skills Journal*, pp. 19 – 39, 2020, [Online]. Available: https://nationalcyberwatchcenter.wildapricot.org/event-4057720. [Accessed: Sept. 30, 2022].

[11] H. Jang, "Identifying 21st century STEM competencies using workplace data," *Journal of Science Education and Technology,* vol. 25, no. 2, pp. 284 – 301, 2016, doi: 10.1007/s10956-015-9593-1.

[12] S.E. Dreyfus, "The Five-Stage Model of Adult Skill Acquisition," *Bulletin of Science, Technology & Society,* vol. 24, no. 3, pp. 177 – 181, 2004, doi: 10.1177/0270467604264992.

[13] US Bureau of Labour Statistics. (2022). *Customer Service Representatives: Occupational Outlook Handbook.* [Online]. Available: https://www.bls.gov/ooh/office-and-administrative-support/customer-service-representatives.htm#tab-4 [Accessed: Oct. 27, 2022].

Editorial Office

Applied Cybersecurity & Internet Governance
Kolska Street 12
01–045 Warsaw, Poland
www.acigjournal.com

Publisher

NASK – National Research Institute
Kolska Street 12
01–045 Warsaw, Poland
www.nask.pl