

Editorial Board

Editor-in-Chief | **Aleksandra Gasztold**

Associate Editor | **Krzysztof Silicki**

Editor | **Dorota Domalewska**

Editor | **Rubén Arcos**

Managing Editor | **Agnieszka Wrońska**

International Editorial Board

Saed Alrabaee

Patrick Burkart

Mu-Yen Chen

Myriam Dunn-Cavelty

Margeret Hall

Marta Harničárová

Joanna Kołodziej

Vijay Kumar Chahar

Salman Ahmed Khan

Sarat Kumar Jena

Mary Manjikian

Andrzej Najgebauer

Eunil Park

Cathryn Peoples

Tim Stevens

Paul Timmers

Jan Valíček

ISSN: 2956-3119

E-ISSN 2956-4395

Information email: contact@acigjournal.pl

Copyright: [Some rights reserved: Publisher NASK](#)



The content of the journal "Applied Cybersecurity & Internet Governance" is circulated on the basis of the Open Access which means free and limitless access to scientific data.

Table of Contents

Letter from the Editor-in-Chief	iv
<i>Aleksandra Gasztold</i>	
The Future of Security Empowerment and the Evolving Methodologies Essential to Counter Rising Threats	1
<i>Mary Ellen Zurko</i>	
Assessing Power and Hierarchy in Cyberspace: An Approach of Power Transition Theory	7
<i>Enescan Lorci</i>	
Ransomware: Why It's Growing and How to Curb Its Growth	38
<i>Joshua Jaffe, Luciano Floridi</i>	
Disjointed Cyber Warfare: Internal Conflicts among Russian Intelligence Agencies.....	65
<i>Cosimo Melella, Francesco Ferazza, Konstantinos Mersinas</i>	
Post-Truth and Information Warfare in their Technological Context	99
<i>Ignas Kalpokas</i>	
Vulnerabilities of Web Applications: Good Practices and New Trends	122
<i>Mateusz Nawrocki, Joanna Kołodziej</i>	
Redefining Systemic Cybersecurity Risk in Interconnected Environments	144
<i>Giacomo Assenza, Alessandro Ortalda, Roberto Setola</i>	
Enhancing Secure Key Management Techniques for Optimised 5G Network Slicing Security	170
<i>Kovid Tiwari, Ajay Kumar Phulre, Devraj Vishnu, Saravanan D</i>	
AI in Disinformation Detection	211
<i>Julia Puczyńska, Youcef Djenouri</i>	
Exploiting Human Trust in Cybersecurity: Which Trust Development Process is Predominant in Phishing Attacks?	233
<i>Morice Daudi</i>	

Denmark's Sector Responsibility Principle: A Tedious Cyber Resilience Strategy	250
<i>Mikkel Storm Jensen</i>	
Vulnerability of Students of Masaryk University to Two Different Types of Phishing	268
<i>Klara Dubovecka</i>	
Deepfake Influence Tactics through the Lens of Cialdini's Principles: Case Studies and the DEEP FRAME Tool Proposal	286
<i>Pawel Zegarow, Ewelina Bartuzi</i>	
An Analysis of Cybersecurity Policy Compliance in Organisations	303
<i>Hugues Hermann Okigui, Johannes Christoffel Cronjé, Errol Roland Francke</i>	
Jobs Exposure to Generative AI: Ongoing Study by NASK-PIB and ILO	322
<i>Marek Troszyński</i>	

Letter from the Editor-in-Chief

Aleksandra Gasztold

Dear Esteemed Readers,

It is with great pleasure that I introduce the 2024 volume of Applied Cybersecurity & Internet Governance (ACIG). From technological advancements to increasingly sophisticated threats, 2024 has been a year of disruption and progress, highlighting the resilience of the cybersecurity community. This issue addresses the critical aspects shaping the current digital ecosystem, with contributions that highlight the breadth and depth of research in cybersecurity studies. The diversity of topics illustrates the complexity of contemporary challenges and the innovation required worldwide to address them.

We begin with an interview with Mary Ellen Zurko's 'The Future of Security Empowerment and the Evolving Methodologies Essential to Counter Rising Threats', which provides a forward-looking perspective on adapting to an ever-changing threat landscape. Similarly, Enescan Lorci's 'Assessing Power and Hierarchy in Cyberspace: An Approach of Power Transition Theory' offers a theoretical framework to understand shifts in global cyber power dynamics. Furthermore, ransomware, a growing and persistent threat, is explored in depth by Joshua Jaffe and Luciano Floridi in 'Ransomware: Why It's Growing and How to Curb Its Growth'. Their work addresses the alarming rise in ransomware attacks and proposes actionable strategies for mitigation. On a related note, Cosimo Melella, Francesco Ferazza and Konstantinos Mersinas examine geopolitical dimensions in 'Disjointed Cyber Warfare: Internal Conflicts among Russian Intelligence Agencies', while Ignas Kalpokas discusses the technological underpinnings of disinformation in 'Post-Truth and Information Warfare in Their Technological Context'.

To address critical cybersecurity challenges, we present the research article by Mateusz Nawrocki and Joanna Kołodziej, 'Vulnerabilities of Web Applications: Good Practices and New Trends' and 'Redefining Systemic Cybersecurity Risk in Interconnected Environments' by Giacomo Assenza, Alessandro Ortalda and Roberto Setola.

Moreover, emerging technologies are also at the forefront of the issue, with 'Optimising 5G Network Slicing with Secure Key Management Techniques' by Kovid Tiwari, Devraj Vishnu and Saravanan D, alongside 'AI in Disinformation Detection' by Julia Puczyńska and Youcef Djenouri. These papers exemplify how innovation can both strengthen and challenge cybersecurity practices. They show the dual nature of technological advancements as both tools for defence and avenues for exploitation.

Broader perspectives on resilience and policy are presented through insightful analyses. The human element in cybersecurity remains critical with contributions such as Morice Daudi's 'Exploiting Human Trust in Cybersecurity: Which Trust Development Process Is Predominant in Phishing Attacks'. We also present regional perspectives. Mikkel Storm Jensen's 'Denmark's Sector Responsibility Principle: A Tedious Cyber Resilience Strategy' examines a national framework for resilience, while Klara Dubovecka's study on phishing vulnerabilities among university students emphasise the importance of understanding human behaviour and trust dynamics. Paweł Zegarow and Ewelina Bartuzi's 'Psychological Analysis of Influence Methods in Deepfake Content and Development of the DEEP FRAME Tool' explores the cognitive vulnerabilities exploited in misinformation campaigns. The study 'An Analysis of Cybersecurity Policy Compliance in Organisations' by Hugues Hermann Okigui, Johannes Christoffel Cronjé, and Errol Roland Francke examines factors influencing organisational compliance with cybersecurity policies, highlighting insider threats, phishing, behavioural resistance and enforcement challenges. These articles provide evidence of the need for continued investment in awareness and education to address human factors in cybersecurity.

We conclude Volume 4 with a commentary, 'Jobs Exposure to Generative AI: Ongoing Study by NASK-PIB and ILO' by Marek Troszyński who examines the impact of generative AI on the labour market and presents an ongoing study by NASK-PIB and the ILO that aims to develop an index to estimate job exposure to AI-driven automation in Poland.

Reflecting on 2024, it is evident that resilience has become a defining theme. For this reason, The European Union (EU) adopted the Cyber Resilience Act (CRA), which establishes uniform cybersecurity requirements for digital products to ensure their security across their entire lifecycle, from design to market. The CRA also addresses gaps in existing regulations and seeks to harmonise cybersecurity measures across the EU. This shift from a prevention-focused

approach to a sustainability-driven strategy has proven essential, as organisations increasingly acknowledge the inevitability of cyber incidents. Preparedness, detection, response, recovery and adaptation are now recognised as critical pillars for maintaining operational continuity and security in an interconnected world.

Looking ahead to 2025, emerging challenges such as quantum computing, generative AI and increasingly complex cyber ecosystems will demand innovative solutions and collaborative efforts. The insights and research shared in this issue will undoubtedly inform and inspire the cybersecurity community in addressing these future challenges.

I extend my gratitude to our authors, reviewers and readers for their dedication and support. Together, we continue to advance the field and strengthen the foundations of cybersecurity.

Wishing you a (cyber) secure 2025,

Aleksandra Gasztold
Editor-in-Chief

The Future of Security Empowerment and the Evolving Methodologies Essential to Counter Rising Threats

Mary Ellen Zurko | MIT Lincoln Laboratory, United States |
ORCID: 0000-0002-9427-1607

Prof. Mary Ellen Zurko was interviewed by Prof. Aleksandra Gasztold online on 4 December 2024

— Drawing from your extensive experience in usable security, what lessons can be applied to the fight against disinformation?

For starters, thank you for inviting me to share my insights on the future of security empowerment and evolving methodologies to counter those threats. While I'm a technologist with CS degrees from MIT, how humans interact with technology has been a continuing interest of mine, going back to my bachelor's thesis (I won't name the year!), and my first job in cybersecurity, where I owned the UI (because no one else wanted to).

I've been active in the usable security research community since defining the area in 1996 (no one had put a name to it, though several others were doing it). One observation from my work in cybersecurity and usable security is that the same patterns and lessons recur (although it is hard to predict WHICH will recur when things change).

Also, I tend to be careful about using the terms disinformation, misinformation, and mal-information. They have different definitions, at least in the research community, that have to do with the intention of the source and sender. Misinformation is the term that

Received: 09.12.2024

Accepted: 10.12.2024

Published: 27.12.2024

Cite this article as:

Zurko, M. E. "The Future of Security Empowerment and the Evolving Methodologies Essential to Counter Rising Threats," ACIG, vol. 3, no. 2, 2024, pp. 1–6. DOI: 10.60097/ACIG/199341

Corresponding author:

Mary Ellen Zurko, MIT Lincoln Laboratory, United States; E-mail: maryellen.zurko@ll.mit.edu

 0000-0002-9427-1607

Copyright:

Some rights reserved

(CC-BY):

Mary Ellen Zurko

Publisher NASK



assumes the fewest ill intentions, and again, because I'm a technologist, I use that one when talking in generalities that do not involve known ill intentions on the part of all sources.

So, actually back to your question. I would say most generally one of the lessons from usable security is that technologists often struggle to predict how humans will react to new technologies, including misinformation. Early research in usable security found pretty rapidly that we had to actually test or otherwise measure how people would respond to new technology, its new uses, and the threats it imposes on them. Not entirely realistic, very controlled in-lab testing would yield responses different from measuring what people do "in the field" (that's what we call "in real life").

How can effective warnings be designed to combat disinformation, and what lessons can be drawn from Facebook's approach to addressing misinformation and user engagement?

So I want to say my first lesson on misinformation was well before college (again, not naming the decade), when I saw an elderly relative regularly reading a newspaper I had never heard of, called the Weekly World News. This was a tabloid of mostly fictional "news", whose most memorable headline was "Bat Child Found in Cave!". The stories were all largely impossible, but explained in terms that made them sound both plausible and quite sensational. She paid money for that newspaper, and we were not people who had a lot of money for unnecessary items. Thus my first lesson, people will actually go out of their way and pay good money to consume misinformation.

Facebook's initial response to misinformation was to identify it. That didn't make Nana avoid it, and it didn't do much for the Facebook population either. It's said that in some cases it attracted people instead. And again, it certainly attracted my Nana. The usable security community has a very long background in researching how users respond to security warnings. I would say one of the takeaways from that is that if the security warnings themselves are equivocal, if they can't be certain and clear about the harm, if it's only the vague possibility of harm, users will click through them, at an increasing rate as they get accustomed to seeing them time and again. So Facebook changed their initial approach from a vague warning, to a "disputed" flag, with pointers to related articles, which countered or debunked the misinformation [1, p. 4]. I'd also say they have the luxury to test "in the field" and at scale, and that was what they were doing.

——— **In your work, you emphasize the importance of layered defenses, which involve applying a combination of strategies to create a comprehensive system for countering information manipulation. Could you elaborate on this approach, including its technical, educational and warning layers?**

Yes, thank you. I'm not only a researcher; I've worked on product. I was security architect for one of IBM's first cloud products. One critical lesson I learned is that securing a system requires viewing it holistically, as a system with all kinds of layered defenses. Layering technical defenses, called "defense in depth", is considered best practice. When humans are involved, making choices and getting things done, then those layered defenses need to include the human, but not get in the way of what they are doing. The evolution of anti-phishing defenses is a great example. Technology alone can't be sure that an email is phishing (or worse still, targeted spear phishing). Various technical responses are unsatisfactory alone, since they can't be sure. Even with anti-phishing education, both the technology and the human can be tricked by ever evolving attacks. Some percentage of users will fall for a strong targeted phishing attack, even when technology, education, and warnings have done their level best (in part because they were also doing their level best on email that wasn't phishing). So the system needs to be designed not only to defend against threats but also to anticipate and mitigate the impact of inevitable breaches and breakdowns.

——— **You and your team have developed the CIOTER system, which integrates these principles into a robust and scalable testbed.**

Yes, thank you. Sorry to interrupt you mid point! I have a passionate belief in the importance of testing, in both cybersecurity and usable security. So the goal of developing a testbed for Countering Influence Operations (the CIO in 'seaoatter'), by testing the technology involved and the human use of that technology is an exciting one for me.

——— **Can you explain the CIOTER system that you and your team developed. Can you present its purpose, design principles, and potential applications in evaluating and advancing tools for information operations?**

I'll try to keep it crisp, but any reader interested in all the wonderful details can read our published paper "A Testbed for Operations in the Information Environment" [2].

CIOTER focuses on building testbed capabilities for assessing technology used in Operations in the Information Environment; technology used to detect and counter misinformation and its cousins. It is inspired in part by cybersecurity testbeds, which are used extensively in education, technology training, and exercises in cybersecurity skills. While cybersecurity testbeds largely focus on network and host based attacks and defenses, our OIE focus is on testbed capabilities that focus on human-readable data and content, services like social media, and how human operators can work with tools to detect and counter misinformation.

We've designed our capabilities to be reusable, redeployable, and reconfigurable, so that they can be used in a variety of contexts, and can interoperate with and complement cybersecurity testbeds.

How does CIOTER's modular architecture facilitate the integration of emerging technologies or adaptation to new adversarial tactics in information operations?

From a technical infrastructure perspective, CIOTER's modularity is achieved through containerization, allowing mix and match with different technologies that process content that might include misinformation in any format; text, memes, videos. A significant focus of CIOTER is on the content pipeline, which not only processes information but also archives and curates different datasets that can represent different adversarial tactics and technologies over time. We can even iteratively test technologies that generate and detect technical changes in content, such as modifications using different forms of AI or ML [3].

In the context of combating disinformation, how does CIOTER contribute to the development of tools like deepfake detection or authorship verification systems?

Both deepfake detection and authorship verification are fairly mature uses of AI technology to detect misinformation. There are curated datasets available, and competitions with established metrics for how well a piece of technology does over a specific dataset. CIOTER can be used to try out a new technology, or an established technology over a dataset modified with a new or different approach. Our extensible metrics engine has all the accepted metrics for success of AI on these tasks, and can be modified with new ones that are tuned to different tradeoffs in things like false warnings. For example, we compared the performance of a specific deepfake detection approach over a corpus that included AI generated

deepfakes, and another dataset representing a different threat model; manually modified images (sometimes called “cheap fakes”).

—— **Which aspects of disinformation are most easily analyzed using CIOTER? Does the system allow for the evaluation of the effectiveness of counter-disinformation campaigns?**

We’ve got a cool “Over The Shoulder” capability that lets training organizers see how learners and operators are using technology for countering disinformation and other forms of adversarial content. It records all the interactions for viewing during training, and analysis after the event. If a student is confused, or something goes wrong with the tool or how they used it, instructors can replay the session to pinpoint the issue and help. If there was a ‘right’ answer and the learner didn’t identify it, graders can use the recording to give partial credit if the right keywords appeared, for example, by searching for them. CIOTER also includes dashboards that can show all kinds of activity during an event, for one participant, or a team. The measurements can be correlated with demographic information, so you can look at how different experience levels or roles influence tool use and task completion.

—— **Given the rapid evolution of social media platforms and adversarial techniques, how does CIOTER remain agile and relevant in addressing new threats?**

One thing we all know is that social media platforms will come and go, evolve and change. The specific features at a point in time on a social media platform will mean different things at different times (like the blue check on Twitter accounts), and “the algorithm”, which determines what each individual sees, will change and effect the impact of both adversarial content and counter disinformation content. To address this, CIOTER is designed to remain agile by incorporating a capability that can flexibly emulate a specific social media platform at a specific point in time, to allow for replay of curated datasets and generated content that reflects what it looks like in various platforms, under different, configurable assumptions.

—— **What do you see as the most critical areas for future research in countering influence operations? Are there specific technological or interdisciplinary advancements you believe are essential to developing more effective defenses against disinformation?**

One lesson I learned from pioneering usable security is how challenging it can be to publish research that crosses

established boundaries in existing conferences and journals. Rising PhDs and professors need to get their research published, so need to work in areas that are publishable. I've heard professors say that their research on countering influence operations can suffer from this problem; a cybersecurity venue might think it's sociological research, and a sociological venue might point back to cybersecurity publishing opportunities. Just focusing on cybersecurity problems, I'm on a National Academies study of Cyber Hard Problems, and recorded public testimony available on the website includes discussion of how many cybersecurity problems today go beyond just technical problems.

Defending against disinformation and mal-information can involve not just cybersecurity and sociology, but psychology and even political science. There aren't a lot of venues that have specialist reviewers in all those areas. Fostering the best research in countering malign influence operations will require building those communities and venues, that support interdisciplinary work.

Mary Ellen Zurko is a technical staff member in the Cyber Operations and Analysis Technology Group at MIT Lincoln Laboratory. With over 35 years of experience in cybersecurity and more than 20 patents, she defined the field of user-centered security in 1996. Zurko has worked in research, product development, and early prototyping, and was the security architect of one of IBM's first cloud products. She is a founding member of the National Academies' Forum on Cyber Resilience and serves as a Distinguished Expert for the National Security Agency's Best Scientific Cybersecurity Research Paper competition. Her areas of research focus on unusable security for attackers, zero trust architectures for government systems, security development and code security, authorization policies, high-assurance virtual machine monitors, the web, and public key infrastructure.

References

- [1] M. E. Zurko, "Disinformation and Reflections From Usable Security," IEEE Security & Privacy, vol. 20, no. 3, pp. 4-7, 2022, doi: [10.1109/MSEC.2022.3159405](https://doi.org/10.1109/MSEC.2022.3159405).
- [2] A. Tse, S. Vattam, V. Ercolani, D. Stetson, M.E. Zurko, "A Testbed for Operations in the Information Environment," CSET '24: Proceedings of the 17th Cyber Security Experimentation and Test Workshop, pp. 83-90, 2024, doi: [10.1145/3675741.3675751](https://doi.org/10.1145/3675741.3675751).
- [3] M. E. Zurko, J. Haney, "Usable Security and Privacy for Security and Privacy Workers," IEEE Security & Privacy, vol. 21, no. 1, pp. 8-10, 2022, doi: [10.1109/MSEC.2022.3221855](https://doi.org/10.1109/MSEC.2022.3221855).

Assessing Power and Hierarchy in Cyberspace: An Approach of Power Transition Theory

Enescan Lorci | College of Social Science, Institute of China and Asia-Pacific Studies, Taiwan | ORCID: 0000-0003-0111-6331

Abstract

This study explores the application of Power Transition Theory (PTT) to cyberspace, aiming to establish a comprehensive framework for understanding and measuring cyber power. Utilizing PTT's national power model, the research treats states as rational and unitary actors, integrating the rational actor model to assess state behavior in cyberspace. The objectives include defining cyber power, developing a novel metric for its evaluation, and categorizing states within a hierarchical structure of cyber power. By analyzing key components such as data resources, digital economic strength, and cyber political capacity, the study provides a nuanced understanding of cyber power dynamics. The results demonstrate that the traditional IR theories retain relevance in the cyber domain, offering a valuable lens for comprehending global cyber governance and geopolitical competition. This foundational work sets the stage for future analyses of power transitions within cyberspace, highlighting the critical interplay between traditional power metrics and emerging digital landscapes.

Keywords

power transition theory, cyber power, power assessment, internet population, digital economy, cyber political capacity

Received: 25.03.2024

Accepted: 20.06.2024

Published: 27.07.2024

Cite this article as:

E. Lorci "Assessing Power and Hierarchy in Cyberspace: An Approach of Power Transition Theory," ACIG, vol. 3, no. 2, 2024, pp. 7–37. DOI: 10.60097/ACIG/190481

Corresponding author:

Enescan Lorci, College of Social Science, Institute of China and Asia-Pacific Studies, Taiwan; E-mail: enescanlorci@g-mail.nsysu.edu.tw

 0000-0003-0111-6331

Copyright:

Some rights reserved:
Publisher NASK



1. Introduction

Following the inception of the internet, American policy-makers recognized its potential significance not only in matters of security but also in terms of its economic and ideological impact [1, p. 78]. This critical role of the internet became particularly evident during the Clinton presidency, prompting the American government to take measures aimed at regulating the advancement and dissemination of internet-related technologies. In response to the increasing involvement of governments in cyberspace during this period, John Perry Barlow, a cyberlibertarian activist, composed his renowned “Declaration of Independence of Cyberspace” in 1996 [2]. In this declaration, Barlow contended that cyberspace should remain free from the interference of governmental entities, asserting that it is not a domain amenable to the practice of sovereignty by governments from the industrialized world.

Barlow’s perspective emphasized that governments should not exercise hegemonic control over cyberspace. Despite the establishment of the Internet Corporation for Assigned Names and Numbers (ICANN) in 1998, which granted the American government a significant degree of influence over cyberspace, the ideals of a free and open cyberspace, as well as the unimpeded flow of information, were conducive to furthering American objectives of disseminating liberal economic principles and democratic values worldwide in the post-Cold War era [3].

In this context, the United States has significantly influenced the progression of the internet and other Information and Communication Technologies (ICTs) within the global landscape, including cyberspace. Consequently, the governance of this emerging domain has been executed through a model that aligns with American objectives, referred to as the “Multilateral Governance” model [4]. Under this model, decision-making authority over cyberspace is shared among governments, non-governmental organizations, private technology companies, and individual actors, all of whom influence the governance framework.

Nonetheless, over time, geopolitical dynamics have resurfaced and begun to extend their reach into the cyber realm, transforming cyberspace into a newfound arena for competition and power politics among major global powers [5]. The escalating dissatisfaction expressed by China and Russia concerning the existing structure of cyberspace, coupled with the United States’ aspiration to uphold its longstanding dominance in this domain, underscores the evident role of power politics in shaping 21st-century great power

competition over cyberspace [6]. In any competition, determining each participant's position necessitates applying suitable metrics, which also holds true for the context of great power competition within the cyberspace realm. Assessing a state's relative standing vis-à-vis others requires evaluating its power capacity to compete effectively in cyberspace.

Assessing national power is a well-established endeavor, traditionally relying on metrics such as economic size, military prowess, population, and other tangible indicators. Nevertheless, the evaluation of cyber power presents a distinctive and intricate challenge [7]. Unlike conventional measures of national power, gauging cyber power is a relatively novel and arduous task [8]. Addressing this complexity necessitates an initial endeavor to precisely define power within cyberspace. Only upon establishing a clear conceptual framework for cyber power can a viable model for its measurement be formulated [9].

Following cyberspace's discernible impact on world politics, scholars of international relations have displayed varied reactions. Some scholars have avoided incorporating cyberspace into their studies, relegating it to low politics. Others have perceived cyberspace as a novel domain that defies the application of traditional international relations (IR) theories. In contrast, many international relations scholars have asserted that traditional IR theories can retain their relevance within cyberspace and have endeavored to apply them to the cyber domain.

For instance, Beltz and Steven adopted Barnett and Duvall's taxonomy for national power and adapted it to cyberspace to conceptualize cyberpower [10, p. 33]. Similarly, Joseph Nye extended his notions of hard and soft power to the context of cyber power [11]. Despite their differing approaches, these scholars shared a common belief in the potential utility of traditional IR theories within cyberspace. They contend that such theories can serve as valuable starting points in comprehending this emerging domain and the competitive dynamics that unfold within it.

In alignment with the abovementioned perspective, this research also subscribes to the notion that IR theories remain relevant and applicable in cyberspace. Embracing this belief, the study applies the Power Transition Theory (PTT) to cyberspace, aiming to achieve several objectives. Firstly, the research endeavors to define cyber power, offering a novel metric for its assessment akin to the model presented by PTT for evaluating national power. Moreover,

beyond proposing a model for measuring cyber power, the study introduces a novel categorization scheme for states in cyberspace, classifying them into four distinct categories: global cyber leaders, cyber great powers, cyber-dependent powers, and non-cyber powers. This classification serves as a valuable contribution to the field, providing a nuanced understanding of the differing positions and roles assumed by states within the cyber domain.

It is essential to clarify that the study does not address the concept of “power transition” within cyberspace at this initial stage. Instead, its primary aim is to define and propose a novel measurement method for national cyber power, thereby positioning states within a hierarchical order. By borrowing PTT’s national power definition and measurement model, this research establishes the relevance of traditional IR theories to the cyber domain. This foundational work is crucial as it sets the stage for future analyses of power transitions in cyberspace, which can only be thoroughly examined once cyber power has been accurately measured using the proposed model.

Critics might argue that applying PTT to cyberspace without directly exploring “transition” dynamics is premature. However, this study is a preliminary effort to introduce an IR perspective on the definition and measurement of cyber power. It lays the groundwork for future research. The proposed model must be utilized to measure national cyber power comprehensively and subsequently explore the dynamics of power transitions within this context.

By applying traditional IR theory to cyberspace and demonstrating its applicability, this research addresses a significant gap in the existing literature. It also puts forward an innovative model for measuring cyber power and provides valuable insights into the hierarchical structure of states within cyberspace. These substantial contributions offer a new lens to understand global cyber governance and geopolitical relations in this emerging and critical domain.

The study is organized as follows to achieve these objectives. First, it reviews the literature on definitions and measurements of cyber power, situating the research within existing scholarship and highlighting its contributions. Next, it discusses the hierarchical categorization of countries in cyberspace. Subsequently, the study provides a detailed exposition of PTT’s national power model, elucidating its theoretical framework and approach to defining and assessing cyber power. This structured approach contributes to the international discourse on cyber power and sheds light on the

ongoing competition for cyber dominance among major global actors. By offering a comprehensive understanding of cyber power, its dynamics, and its impacts on global affairs, this research aims to inform policymakers about the implications of their actions in cyberspace, ultimately striving to create a safer cyber environment for all stakeholders.

2. Intersection of Cyberspace, Power, and International Relations

According to Nye, the concept of power lacks a universally accepted definition and remains subject to contestation, with individual interpretations reflecting one's interests and values [11]. For instance, realist scholars in International Relations emphasize military power as a cornerstone of national power [12]. On the other hand, liberal perspectives on power encompass a broader spectrum, encompassing non-coercive means to achieve desired outcomes. In the constructivist framework, power is viewed as a socially constructed phenomenon influenced by prevailing ideas, norms, and identities. Here, power extends beyond material capabilities, encompassing the capacity to shape and influence the prevailing norms and values that inform state behavior [13].

Despite the various descriptions of power put forth by different schools of thought in the discipline, a common thread prevails: power is widely acknowledged as a crucial instrument for achieving desired outcomes in international politics [14]. The quantification of power has become a central concern for states, as it enables the assessment of the feasibility and effectiveness of particular actions. States with greater power are likelier to advance their objectives than weaker states. Consequently, power measurement has garnered significant academic attention, mirroring the importance accorded to power and facilitating comparative assessments between different actors, which has become a pivotal activity for decision-makers. When evaluating national power, numerous factors are considered, including territory, wealth, military strength, armies, navies, and military arsenals. These tangible indicators provide insight into a state's potential and capacity to exert influence in the international arena [7].

Unlike material power, the notion of power and its quantification in cyberspace has emerged as a relatively recent focus of academic inquiry. Inkster underscores the significance of assessing power by contending that the absence of reliable metrics could lead to mission failure [8]. States must gauge their power and that of their

adversaries to ensure their security. However, the intricacies of assessing power in cyberspace necessitate a preliminary explication of the concept itself. Without a comprehensive understanding of cyber power, any measurement strategy would prove ineffective [15]. Consequently, a clear and nuanced description of cyber power becomes a foundational prerequisite for developing an effective and meaningful approach to measuring it.

The involvement of IR scholarship in cyberspace dates back to the late 1990s and early 2000s, when the internet and related technologies began to play a crucial role in national security, the economy, and foreign policy objectives. Consequently, a significant body of IR literature has emerged, focusing on cyberspace, cyber power, and cyber warfare from offensive and defensive perspectives.

One of the early seminal works in this field is by Arquilla, who discussed the concept of cyber war and its potential impact on future conflicts, highlighting the strategic significance of cyberspace in international relations [16]. Martin C. Libicki analyzed how control over information can influence the battlefield, affect decision-making processes, and disrupt adversaries' operations. He emphasized the importance of protecting one's information infrastructure while targeting and exploiting vulnerabilities in opponents' systems [17].

Manuel Castells introduced the concept of the network society, where digital networks significantly shape power dynamics and international relations [18]. Similarly, Saskia Sassen explored how globalization and digital technologies influence state sovereignty and global governance, providing foundational insights into understanding cyber power [19].

Keohane and Nye examined how the information age transforms power structures and interdependence among states. Their work laid the groundwork for understanding cyber power in IR [20]. In *"Information Technologies and Global Politics: The Changing Scope of Power and Governance,"* Rosenau and Singh explored how power is redefined in the context of information technologies [21]. They argued that cyber power encompasses control over IT infrastructure, cyber capabilities, and the ability to influence information flows, extending beyond traditional state-centric views and recognizing the significant roles played by non-state actors.

Nissenbaum integrated ethical considerations with IR theories to discuss the implications of cybersecurity for national and international security, highlighting the moral and strategic dimensions of cyber

power [22]. Chadwick, in *"Internet Politics: States, Citizens, and New Communication Technologies,"* explored how the internet and digital communication technologies influence political power and state-citizen interactions, which are relevant to IR and cyber power [23].

Deibert and Rohozinski analyzed how state actors exert control over cyberspace and the impact of these actions on international security dynamics. Their work integrates concepts from IR theories, such as realism and constructivism, to explain the strategic behavior of states in the digital realm [24]. Choucri, in *"Cyberpolitics in International Relations,"* provided a comprehensive examination of how cyberspace intersects with traditional IR theories, discussing how concepts like power, sovereignty, and interdependence are redefined in the context of global cyberspace [25].

Jon R. Lindsay examined the Stuxnet cyber-attack through the lens of IR theory, mainly focusing on deterrence and coercion. He argued that traditional concepts of military power and strategy apply to understanding cyber operations and their impact on international relations [26]. Nye posited that cyber power entails the capacity to achieve desired outcomes by leveraging electronically interconnected information resources within the cyber domain. Conversely, Armistead focused on the role of information in describing cyber power, defining it as the control over a greater volume of information (data) relative to other actors [15].

Eventually, although these diverse perspectives reflected the evolving and multifaceted nature of cyber power and underscored the complexities involved in defining and understanding this concept within the context of cyberspace, it is essential to acknowledge their limitations because, in many of these approaches, the authors see developments in cyberspace either from a defensive or offensive perspective. However, this study argues that defining cyber power solely based on defensive or offensive cyber capabilities may lead to erroneous assessments, rendering assessment of cyber power inconsequential [9].

Instead, a more comprehensive approach is necessary, wherein a cyber-capable state exhibits proficiency in safeguarding the integrity of its cyberspace through vigilant monitoring, timely patching, and proficient network system definition. In addition to defensive capabilities, a cyber-capable state must demonstrate the capacity to manage, manipulate, and effectively navigate vast volumes of data crucial for modern economies and networked military operations [27]. The ability to generate intelligence and strategically wield

cyberspace to exert influence is also imperative. In sum, any definition of national cyber power ought to adopt a holistic approach, considering all facets of the cyber domain beyond mere considerations of defense or offense [27]. Embracing this comprehensive perspective will enable a more accurate and insightful assessment of cyber power, avoiding oversimplifications and yielding more meaningful results in cyber power measurement and analysis.

In this context, this study tries to adopt a holistic approach to understanding cyber power and its assessment, which aims to achieve this by focusing on the objectives pursued by a country within cyberspace. This perspective is in line with the insights provided by scholars like Armistead, who underscore the significance of considering the “context” when defining power [15]. Similarly, Nye argues that the statement of “actor A has power” lacks substantial meaning without specifying the specific scope or purpose for which that power is wielded, i.e., power “to do what [11].” Hence, in the discourse on power in cyberspace, a pertinent point of departure is to inquire into the objectives states seek to accomplish through their cyber capabilities. A comprehensive understanding of the context in which their power is exercised is established by elucidating the aims and desired outcomes that countries aspire to achieve within cyberspace.

Thus, this study argues that effectively assessing cyber power involves assessing a country’s capacities to actualize the objectives it has set for itself in cyberspace. Such an evaluation yields reliable metrics for gauging a country’s cyber power’s extent and potential to influence and shape outcomes in this dynamic domain.

Assessing cyber power from the perspective of “objectives,” the notion of standardizing the concept of cyber power may face challenges due to the potential variations in objectives in cyberspace among different countries. While it is true that objectives may vary somewhat, it is essential to recognize that many objectives are shared among rational states. Thus, analyzing this issue through the lens of the “rational state” assumption can provide valuable insights. When considering the question, “What would a rational state seek to achieve in cyberspace?” the answers likely exhibit significant commonalities. For this reason, this study adopts the assumption of a “rational state in cyberspace,” which allows for generalizing objectives in cyberspace.

By applying this rational state concept to cyberspace, this study distinguishes itself from prior studies and makes valuable

contributions to ongoing discussions. This approach acknowledges the common ground among rational states regarding their objectives in cyberspace, facilitating a more comprehensive understanding of the factors driving cyber power dynamics. By incorporating the rational state assumption, the study provides a framework that accommodates shared objectives and enables a more cohesive and comparative analysis of cyber power among different states. Consequently, the research offers new perspectives and insights that contribute to advancing knowledge and dialogue on cyber power in the contemporary international arena.

Nonetheless, it is essential to acknowledge that despite sharing rational motivations, some countries might encounter challenges in promptly implementing their intentions or using their capabilities. Such obstacles can arise due to the country's regime type and bureaucratic structure, which may affect the speed and efficiency of decision-making processes. Consequently, the domestic political structure can influence a country's cyber power. Thus, it is crucial to consider domestic factors when assessing state cyber power.

In this particular context and under the rational state assumption, this research focuses on three critical objectives related to cyber power. These objectives are pivotal for a rational state striving to secure cyberspace and advance its interests in this domain. To measure a country's capabilities in achieving these objectives, the study employs a set of 30 domestic and international indicators, which serve as evaluative criteria (see Table 1).

Before introducing these capabilities, it is important to emphasize that the model presented in this research is rooted in the PTT's state power assessment strategies. Therefore, it is essential to provide a concise overview of the PTT's key principles and concepts, especially in regard to power. Subsequently, the study will proceed to apply the PTT's power assessment framework within the context of cyberspace, first by introducing hierarchical situations in international cyberspace and then introducing a model for assessing a state's cyber power.

3. PTT's Approach to National Power and International Hierarchy

The Power Transition Theory (PTT) 's central premise revolves around significant power shifts within the international system, leading to periods of either stability or conflict. These shifts are often characterized by the ascent of a challenger power

Table 1. Synthesized model for cyber power assessment.

Three vital objectives of the rational state in cyberspace	Indicators for the assessment of capabilities	
	Domestically	Internationally
Attainment of a substantial internet population and ownership of data	Effectiveness of domestic cyber intelligence	Effectiveness of international cyber intelligence
	Effectiveness of domestic cyber surveillance	Effectiveness of international cyber surveillance
	Effectiveness of domestic offensive cyber operations	Effectiveness of international offensive cyber operations
	Effectiveness of domestic defensive cyber operations	Effectiveness of international defensive cyber operations
	Effectiveness of domestic cyber influence operations	Effectiveness of international cyber influence operations
Cultivation of a robust digital economy	Amount of domestic broadband infrastructure (ICTs, Internet, and IT(data) Sectors), ICT employment	Amount of international broadband infrastructure (ICTs, internet infrastructure, 5G, AI,IT)
	Level of domestic e-commerce sales	Level of international e-commerce sales
	Domestic digital payment adoption	International digital payment adoption
	Share of ICTs in total GDP And ICT access and use by households and individuals,	Share of ICT exports in the country's overall export
	Effectiveness of digital government services	Digital economic trade agreements
Cultivation of a high degree of cyber political capacity	Effectiveness of capacity building and awareness	Capability of determining international cyber norms, principles, standards, and developments (International cyber governance)
	Capability of making effective National cybersecurity strategies	International treaties and agreements
	Capability of making and implementing Cybersecurity Laws and Regulations	Participation in International Fora
	Capability of data gathering protection and privacy	Participation in Cybersecurity Cooperation Agreements
	Quick and effective incident response and coordination	Active cyber public diplomacy

that challenges the existing dominant power. PTT posits that such power transitions bear substantial consequences for international politics, influencing the potential for conflict or cooperation among states. As power constitutes a major determinant of war and peace in the international system, PTT places great emphasis on explaining its dynamics [28].

PTT conceptualizes national power as a composite of three crucial elements: population, economic productivity, and political capacity [29]. The first element is population, which encompasses the sheer number of people and the quality of human resources. This includes the population's skills, education, health, and demographic characteristics. A robust and skilled population is essential for sustaining economic growth, supporting national defense efforts, and contributing to innovation and technological advancements. A large population provides a substantial labor force necessary for industrial and economic development. It also offers a wide recruitment base for the military, enhancing a nation's defense capabilities. Furthermore, the population's age structure plays a critical role; a younger, dynamic workforce can drive economic productivity, whereas an aging population might strain social services and economic growth [30].

The second element is economic productivity, typically measured by a country's Gross Domestic Product (GDP). Economic productivity reflects a nation's capacity to generate wealth and economic output, which underpins its ability to invest in various sectors critical for national power, such as military capabilities, technological advancements, and infrastructure development [31]. A strong economy enables a country to sustain prolonged periods of conflict by financing military operations, maintaining sophisticated defense systems, and ensuring economic resilience in the face of blockades or sanctions. Economic productivity also enhances a nation's diplomatic leverage as economic aid and trade agreements become tools of influence. Moreover, a thriving economy attracts global investments and fosters innovation, further solidifying a nation's competitive edge in the international arena.

The third element is political capacity, referring to the effectiveness of a country's political system in mobilizing resources from its citizens and deploying them efficiently to achieve national objectives. Political capacity involves the ability of the state to enact policies, maintain internal stability, and project power externally [32]. An efficient political system can harness the potential of a large population and a productive economy by ensuring that resources are directed toward strategic goals. This includes the capability to implement sound economic policies, maintain law and order, provide public goods, and manage crises. Political stability and governance quality are crucial for fostering an environment where economic and human resources can thrive. Political capacity also encompasses the ability to form strategic alliances and exert influence in international institutions. A politically capable state can navigate complex

global challenges, mediate conflicts, and shape international norms and rules to its advantage.

These three components, population, economic productivity, and political capacity are interdependent and mutually reinforcing. A nation may have a large, economically productive population, but its power potential remains constrained without an adept political system to harness and utilize these resources effectively. Conversely, a small nation with a highly efficient political system can maximize its limited resources to achieve significant influence.

For instance, China's rise is often attributed to its large and increasingly skilled population, rapid economic growth, and a political system capable of mobilizing resources for large-scale projects and strategic initiatives. On the other hand, countries with abundant resources but weak political systems, such as some oil-rich states, may struggle to convert their potential into sustained national power. Thus, PTT views power as a product of a country's harmonious domestic components. This holistic approach underscores that national power is not merely a function of economic or military might but also depends on the quality and effectiveness of political institutions and the nation's human capital. Understanding these dynamics is crucial for analyzing power transitions, as shifts in the relative power of states can lead to significant changes in the international order. By examining how national power is constructed and distributed within this framework, PTT provides valuable insights into the stability and transformation of the global system.

On the other hand, in the context of PTT, the distribution of power within the international system is depicted as a hierarchical structure. At the apex of this hierarchy lies the dominant power, which exercises control over a significant portion of the system's resources and sets the rules and norms that govern international interactions. The dominant power acts as the primary architect of the international order, establishing institutions and frameworks that reflect its interests and values [33, p. 86].

Below the dominant power are the great powers, which possess considerable capabilities and resources, though not to the extent of the dominant state. Great powers play significant roles in shaping international politics and can challenge or support the dominant power's leadership. They have substantial military, economic, and political influence, allowing them to impact global governance and security dynamics.

Further down the hierarchy are the middle powers with moderate resources and capabilities. Middle powers often act as stabilizers within the international system, supporting the existing order or advocating for incremental changes. They may form coalitions with other states to amplify their influence and contribute to regional stability and development.

At the bottom of the hierarchy are the minor powers, which possess the fewest resources and capabilities within the system. Small powers are often more vulnerable to external pressures and have limited ability to influence global politics independently. They typically align with more powerful states or international organizations to safeguard their interests and security.

This hierarchical arrangement underscores the varying degrees of influence and authority among states in the international system. The dominant power, with its superior resources, assumes the role of founder, rule-maker, and value determinant of the international system [34]. Meanwhile, great, middle, and minor powers navigate the international landscape based on their respective capabilities and positions within the hierarchy. This structure shapes the interactions between states, influencing the patterns of conflict, cooperation, and competition in global politics. According to Rachel, understanding these dynamics is crucial for analyzing power transitions, as shifts in the relative power of states can lead to significant changes in the international order. By examining how national power is constructed and distributed within this hierarchical framework, PTT provides valuable insights into the stability and transformation of the global system. Following this elucidation of PTT, the subsequent section of this research will establish an international power hierarchy in cyberspace and develop a novel approach to understanding and evaluating cyber power inspired by the foundational principles of PTT (see Figure 1).

4. Hierarchy in Cyberspace

Hierarchy in cyberspace posits that a dominant cyber power occupies the pinnacle, exerting control over most resources in the cyberspace domain. This dominance is characterized by a substantial command over critical infrastructures, advanced technological capabilities, and significant cyber intelligence assets. Importantly, as in Power Transition Theory (PTT), being the dominant cyber power does not necessarily equate to being a hegemon [29]. While a hegemon exercises unrivaled supremacy and exerts influence unilaterally, the dominant cyber power's influence is more nuanced and collaborative.

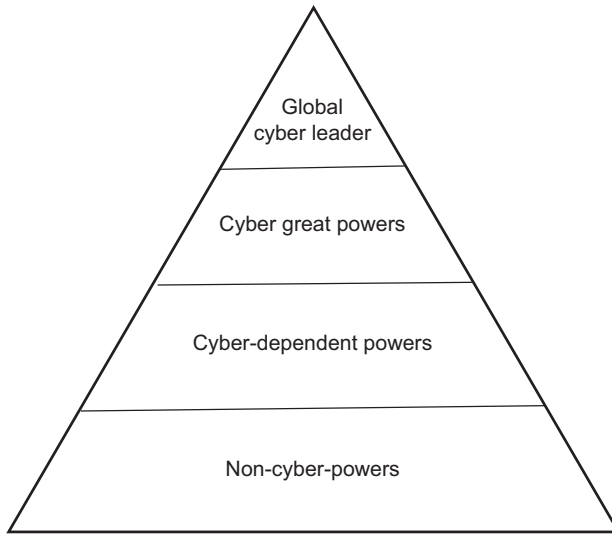


Figure 1. Cyberspace power hierarchy.

Instead, the dominant cyber power assumes a leadership role in advancing technical developments within cyberspace and in shaping the standards, norms, principles, and regulations governing cyberspace. This involves pioneering innovations in cybersecurity, artificial intelligence, and data governance that set benchmarks for others to follow. By establishing frameworks and protocols for secure and efficient cyber operations, the dominant cyber power influences global practices and policies.

Furthermore, the dominant cyber power aligns these standards and norms with its national and allies' interests. This alignment is achieved through diplomatic efforts, international agreements, and active participation in global forums dedicated to Internet governance and cyber norms. By doing so, the dominant cyber power ensures that the regulatory environment of cyberspace reflects its strategic priorities, such as the promotion of a free and open Internet, protection of intellectual property rights, and establishment of robust cybersecurity measures.

In addition to technical and regulatory leadership, the dominant cyber power also plays a crucial role in shaping the geopolitical landscape of cyberspace. This includes leveraging its cyber capabilities to influence global economic activities, conduct cyber espionage, and engage in strategic cyber operations that reinforce its geopolitical objectives. Through such activities, the dominant cyber power can project its influence across borders, affecting the

internal dynamics of other states and steering international relations in favorable directions.

Following the dominant cyber power, the cyber power hierarchy includes several cyber great powers, each wielding substantial influence within cyberspace. The stability and maintenance of the cyberspace system largely depend on the satisfaction of these cyber great powers with the existing framework [35]. Domestically, cyber great powers demonstrate capabilities in data control, possess robust digital economies, and exhibit strong political cyber capacity. However, their willingness to exercise their cyber-political capacity internationally is contingent upon their satisfaction with the prevailing system.

For example, the European Union (EU), a great cyber power with significant capabilities, refrains from challenging the USA to assert its cyber-political capacity internationally. This is primarily due to the existing structure of international cyberspace, which aligns with the EU's national interests by emphasizing freedom, free flow of information, liberal economic principles, and decentralized decision-making processes. In contrast, despite possessing substantial cyber capabilities, including vast data control, robust digital economies, and effective domestic cyber political capacity, other great cyber powers such as China and Russia remain motivated to enhance their cyber political capacity on the international stage [36]. This is driven by their dissatisfaction with the current system, particularly the governance model of cyberspace. Unlike the USA and its Western allies, China and Russia advocate for a more pronounced role of the state in cyberspace and full sovereignty of states in this domain [37]. Consequently, these challengers, having already bolstered their other cyber capabilities, are now earnestly endeavoring to augment their international cyber-political capacity to reshape the USA-led structure of cyberspace.

After the cyber great powers, many cyber-dependent powers are situated within the cyber power hierarchy. These states possess certain cyber capabilities, including a substantial internet population and a degree of digitalization with accessible internet services for their citizens. However, they rely on foreign technologies for critical services such as internet infrastructure, telecommunication technologies, 5G, and AI. Consequently, while cyber-dependent countries have control over some data due to their internet population and digital aspects of their economy, their reliance on external actors to develop these capabilities renders them vulnerable in terms of cyber security. This vulnerability is particularly evident in examples of cyber-dependent powers in many developing nations.

Finally, at the bottom of the international cyberspace hierarchy, we find numerous non-cyber powers characterized by the absence of a fundamental component and source of cyber power, namely the internet population. These states exhibit an internet penetration rate of less than fifty percent and face the challenge of expensive internet services beyond the affordability of their citizens. The World Wide Web Foundation, Alliance for Affordable Internet employs a metric indicating that for internet services to be deemed affordable, 1GB of data should cost 2% or less of the average monthly income [38]. However, numerous African countries fail to meet this criterion, as exemplified by Equatorial Guinea, where 1GB of mobile data costs a significant \$49.67 [38]. Sao Tome Principe and Malawi follow closely with costs of \$30.97 and \$25.46 per gigabyte, respectively. In addition, Chad and Namibia are in the top five, with average prices of \$23.33 and \$22.37 per gigabyte, respectively. These circumstances underscore the challenges faced by non-cyber powers in attaining affordable and accessible internet services, limiting their capacity to partake in the realms of cyberspace and cyber power.

This hierarchical model of cyber power provides a nuanced understanding of states' various roles and capacities within the international cyberspace arena, reflecting the complexity and multifaceted nature of power dynamics in the digital age. By categorizing states into different tiers based on their cyber capabilities, this model elucidates the diverse ways in which states engage with and exert influence in cyberspace

5. Translating PTT's National Power Model to Cyberspace

During the 1950s, the PTT emerged as a distinct theoretical framework, offering a critical perspective on the prevailing balance of power theory. PTT's foundational arguments rest upon key assumptions, notably asserting that states represent the primary units of analysis in the international system and that they act as rational entities in their interactions [39]. The primacy of states as central actors in the international realm found significant acceptance among various international relations theories in the physical domain. However, the applicability of such assumptions encountered challenges when applied to the context of cyberspace.

The distinctive nature of cyberspace complicates the traditional state-centric perspective endorsed by the PTT. Notably, the diffusion of power in the cyber domain transcends the conventional state-centric paradigm, as multiple actors assume prominent

roles alongside states. This includes private companies, endowed with substantial capabilities in the cyber realm, and individual actors who significantly influence and shape cyberspace dynamics [39].

Nye's analysis in cyberspace discerns three distinct actor categories: governments, organizations with well-structured networks, and individuals. According to Nye it is true that the power diffused among these actors however the distribution of power does not imply a state of equality in capabilities [11]. Governments, due to their possession of substantial resources, wield greater capabilities within the cyber domain. Moreover, the geographical underpinnings of the internet's physical infrastructure, coupled with governments' sovereign authority over territorial spaces, endow location with continued significance as a valuable resource in cyberspace [40]. In addition, geography serves as a basis for governments to exercise legal coercion and control, as a government can exert power extraterritorially if a market is sufficiently extensive [41]. Non-state actors in the cyber realm must safeguard their legal standing and brand reputation, necessitating strong incentives for compliance with local legal structures. This adherence to the established legal framework becomes another resource of power for governments, given their authority in shaping domestic legal systems [4].

Consequently, despite power diffusion in cyberspace, this does not translate to power equalization, as states remain the primary actors with superior resources and capabilities [11]. Thus, this research assumes state as unitary actor in cyberspace and predominantly focuses on states and their cyber objectives, aiming to discern hierarchy in cyberspace and cyber power dynamics.

As expounded previously, the present research endeavors to conceptualize cyber power through the lens of a rational state's objectives in cyberspace. Building upon the PTT's elucidation of population, economic productivity, and political capacity as pivotal constituents of national power, this study posits three primary objectives that a rational state seeks to pursue in the cyber domain. First, attaining a substantial internet population and data ownership means a rational state aims to foster a sizeable and active internet user base within its territorial confines, signifying the penetration and accessibility of cyberspace to its population. Moreover, the possession and control of data resources become a critical objective, as data ownership is a valuable asset, contributing to insights, analytics, and potential competitive advantages [42].

Second, cultivating a robust digital economy is based on the idea that the rational state endeavors to nurture and bolster its digital economy, recognizing the profound economic implications of the cyber realm. A thriving digital economy is indicative of a vibrant ecosystem encompassing electronic commerce, online transactions, digital services, and innovative technology sectors, enhancing economic growth and competitiveness on the global stage [43].

Finally, cultivating a high degree of cyber political capacity argues that a rational state seeks to amass a considerable level of cyber political capacity, denoting its ability to wield influence, enact policies, and control cyber activities domestically and internationally. This capacity encompasses regulatory frameworks, legislative measures, and governance mechanisms aimed at safeguarding cyber interests, ensuring cyber stability, and projecting cyber influence on the global political landscape [44].

This research undertakes an evaluative examination of the objectives stated earlier to comprehensively understand the cyber power of a rational state within cyberspace. For each objective, the research assesses specific indicators of capabilities to measure state cyber power, differentiating between a country's domestic and international capabilities. The rationale behind analyzing indicators separately for domestic and international contexts derives from the theoretical alignment with the PTT in its approach to national power.

Analogous to PTT's premise that a nation's power hinges on its domestic dynamics, this research contends that a country's cyber power is similarly contingent upon its domestic cyber capabilities. Indeed, a country is unlikely to emerge as a significant international cyber power without first establishing a certain degree of domestic cyber power [45]. For instance, exerting considerable control over the international flow of data is improbable without prior adeptness in managing domestic data flows and enhancing corresponding capabilities. Hence, an accurate assessment of the country's cyber power necessitates an analysis of both domestic and international capabilities for each objective.

Next, the research analyzes each objective individually, expounding upon their significance in determining the cyber power of a rational state in cyberspace. By delving into the multifaceted dimensions of each objective, the research endeavors to offer a nuanced comprehension of the interplay between a state's strategic cyber pursuits and its overall cyber power within the dynamic and evolving cyber landscape.

5.1. Attainment of a Substantial Internet Population and Ownership of Data

Government ownership of data holds considerable significance in cyber power and governance. First and foremost, it empowers governments with access to vast information repositories, which can be leveraged for various purposes, including intelligence gathering, law enforcement, and national security initiatives. By exercising data ownership, governments can employ sophisticated data analytics, machine learning, and artificial intelligence techniques to derive valuable insights from the collected information, contributing to informed decision-making and policy formulation [46].

Furthermore, data ownership facilitates the capacity of governments to monitor, supervise, and safeguard their internet populations against cyber threats and malicious activities [47]. Comprehensive datasets enable governments to conduct cyber surveillance, detect potential threats, and respond to cyber incidents promptly and effectively. Moreover, data ownership is closely linked to the protection of critical infrastructure, as governments can employ data-driven risk assessments to bolster the resilience of essential digital systems and networks [47].

As previously expounded in this study, the significance of domestic cyber capabilities is pivotal in positioning a state as a significant actor in the international cyberspace arena. Within this context, internet population and data ownership are crucial in shaping a nation's overall cyber power. These two concepts are intricately linked, as the size and engagement of the internet population directly influence the generation and accumulation of substantial data arising from their online interactions, activities, and behaviors. Drawing upon the PTT's emphasis on population as a fundamental element of national cyber power and recognizing its role as a resource for economic productivity, this research similarly underscores the internet population's value as a reservoir of data.

As the number of individuals accessing the internet and actively participating in online services continues to rise, the volume and diversity of data generated through their digital activities undergo an exponential expansion. This data encompasses a broad spectrum of information, ranging from personal details to digital communications and user behavior patterns [48]. The data, in turn, assumes a critical asset for states seeking to strengthen their cyber power. Through effective data ownership and governance, governments can harness this vast repository of information to gain insights, make informed decisions, and enhance their cyber capabilities.

Comprehensive data ownership derived from the internet population empowers governments with numerous advantages. It facilitates the development and deployment of advanced data analytics, machine learning, and artificial intelligence techniques, empowering states to derive meaningful intelligence and knowledge from this data reservoir [49].

Data assumes significant importance for cyber power and is often likened to the “new currency” or “new oil” in the digital age. It plays a pivotal role in the development of Artificial Intelligence (AI) technologies, as the underlying logic of AI systems relies on vast volumes of data for learning, comprehension, decision-making, and performance enhancement [49]. The abundance of data correlates with reduced errors in AI systems, making data ownership and accessibility crucial for advancing AI services within a country. The possession and utilization of data have substantial implications for a nation’s cyber intelligence, surveillance, and cyber offensive/defensive operations. A country’s ability to own and manage data can significantly impact its cyber capabilities and prowess. By examining a country’s performance in these areas, valuable insights can be gleaned regarding its proficiency or limitations in data collection and ownership [50].

Similarly, in cyber surveillance, data ownership is instrumental in monitoring and detecting potential cyber threats or illicit activities within a country’s digital infrastructure. Surveillance activities heavily rely on data streams to identify suspicious patterns or behaviors, thereby bolstering the nation’s cyber resilience and situational awareness [51].

Regarding cyber offensive and defensive operations, data plays a critical role in enhancing the efficacy of these activities. Governments can leverage data-driven intelligence to formulate offensive cyber operations, targeting specific adversaries or vulnerabilities. On the defensive front, possessing robust data resources allows for proactive measures in fortifying cyber defenses and responding to emerging threats promptly [52, p. 32].

Data ownership is a cornerstone of a country’s cyber power, impacting various facets of its cyber capabilities. Access to vast and diverse datasets fuels the development of AI technologies and strengthens a nation’s cyber intelligence, surveillance, and offensive/defensive operations [10]. Evaluating a country’s performance in these domains provides valuable insights into its ability to collect, manage, and utilize data effectively, ultimately contributing to its overall cyber power and resilience.

Nevertheless, as the PTT contends that while the population, including the internet population, constitutes a vital resource for national power, it is not the sole determinant [39]. Similarly, in cyberspace, although a high internet population and data are essential objectives for a rational state, they do not encompass the entirety of its pursuits. To comprehensively grasp the dynamics of cyber power, examining a country's performance in its digital economy is imperative. Assessing a country's digital economy offers valuable insights into its ability to effectively harness its resources, such as the internet population and data, to attain digital economic competitiveness.

5.2. Cultivation of a Robust Digital Economy

The global economy is undergoing a profound transformation driven by the rapid advancement and widespread adoption of information and communication technologies (ICTs). Notably, the proliferation of digital data over the internet has been accompanied by the rise of significant technologies such as big data analytics, artificial intelligence (AI), cloud computing, and novel business models. The continuous expansion of internet-connected devices and users and the increasing integration of value chains through digital means further underscores the escalating significance of digital data and technologies [53]. Consequently, the ability to access and leverage data effectively, transforming it into digital intelligence, assumes critical importance in determining the competitiveness of states in the contemporary economic landscape.

The ongoing digitalization process in the global context has led to the emergence of the digital economy, which, at its nascent stage, lacks a universally accepted definition. In the late 1990s, initial analyses of the digital economy primarily centered on the adoption of the Internet and its economic implications [54]. As Internet usage continued to expand, subsequent reports from the mid-2000s onward examined the factors that could facilitate the growth and development of the internet economy.

The digital economy can be defined as a subset of the overall economic output that stems from the utilization of digital technologies and is structured around business models primarily centered on digital goods or services [43]. However, other scholars present a more comprehensive perspective, considering the digital economy as the total economic output derived from diverse digital elements.

These digital inputs encompass various aspects, including digital skills, equipment, digital goods, ICT exports, and digital services

utilized in production. This broader definition allows for a more comprehensive examination of a country's digital economy, whether in the context of its domestic or international dimensions. By analyzing a nation's performance across these digital inputs, valuable insights can be gleaned regarding the economic output (digital economy) generated from these digital resources [43].

For several significant reasons, the digital economy plays a pivotal role in shaping a country's cyber power. First, it serves as a driving force behind technological advancements and innovations in cybersecurity and cyber technologies [55]. The continuous growth of the digital economy necessitates the development of sophisticated cybersecurity capabilities, including robust threat detection and incident response systems.

Secondly, establishing a strong digital economy requires the implementation of sophisticated cyberinfrastructure that supports various cyber operations and services. This infrastructure forms the foundation for effective cyber governance and management [53].

A flourishing digital economy enhances a country's economic competitiveness and global influence in the cyber domain. A strong presence in the digital economy elevates a nation's reputation and standing in the international cyber landscape.

Overall, a thriving digital economy serves as the backbone of a nation's cyber strength and resilience, enabling it to effectively navigate the complexities and challenges of the cyber domain [55].

5.3. Cultivation of a High Degree of Cyber Political Capacity

The objectives pertinent to a rational state's interests in cyberspace encompass data ownership, information management, cybersecurity, offensive capabilities, cyberinfrastructure, and economic aspects of cyber power. However, a comprehensive analysis of cyber power requires the consideration of additional dimensions. Analogous to the Power Transition Theory's emphasis on political capacity as the government's ability to effectively mobilize resources and achieve national objectives, the realm of cyberspace also demands a high degree of cyberpolitical capacity [39].

Cyber-political capacity in cyberspace pertains to a state's capability to wield cyber resources and technologies to achieve its strategic goals and policy objectives. This includes the effective governance and management of cyber operations, cyber policies, and cyber

strategies at the national level. States with robust cyber-political capacities can leverage their cyber capabilities to assert their interests, influence the global cyber landscape, and safeguard their national security in cyberspace.

Domestically, cyber political capacity involves a country's ability to promptly and effectively formulate policy decisions on cyber-related matters. This includes establishing comprehensive cyber strategies, laws, and regulations that optimize the use of cyber resources for advancing national interests in the international cyberspace domain. A state's ability to effectively govern its cyber activities is fundamental to its capacity to project power internationally. Strong domestic cyber political capacity ensures that the state's cyber infrastructure is resilient, its policies are forward-thinking, and its workforce is skilled and adaptable to emerging cyber threats and opportunities. This internal governance forms the backbone of a country's overall cyber power, enabling it to respond rapidly and efficiently to cyber challenges.

On the international stage, cyber political capacity extends to a country's ability to influence the formulation of global cyber norms, regulations, principles, and standards that align with national interests. This aspect of cyber capacity is closely related to the broader concept of international cyber governance. A nation with significant international cyber-political capacity can shape the international cyber domain's rules, thereby exerting influence over how cyberspace is used, regulated, and protected. Effective participation in international cyber policymaking forums, alliances, and coalitions is crucial. Countries with strong international cyber-political capacities can push for norms and regulations that favor their strategic interests, promote global stability, and prevent cyber conflicts.

The importance of cyber-political capacity cannot be overstated. This capacity is a crucial enabler for a state to achieve and maintain cyber power. Cyber political capacity encompasses the strategic governance and management of a state's cyber resources, policies, and operations, aligning them with national objectives. Without effective governance and strategic management, even states with significant data resources and a robust digital economy may find their cyber power potential constrained. A lack of coherent strategy can lead to disjointed efforts, inefficiencies, and vulnerabilities that adversaries could exploit.

Conversely, states with robust cyber-political capacities can maximize the utility of their cyber assets. Effective governance ensures

that cyber activities are coherent, well-coordinated, and strategically aligned with national objectives. This alignment facilitates the seamless integration of cyber capabilities into broader national security and economic strategies, amplifying the impact of cyber initiatives. For instance, comprehensive cyber strategies can enhance defensive measures against cyber threats, ensure critical infrastructure protection, and bolster the state's ability to conduct offensive cyber operations when necessary.

Moreover, the ability to shape international cyber norms and policies in an interconnected world provides a strategic advantage. States with substantial cyber political capacity can actively participate in international forums, influence the development of global cyber norms, and advocate for policies that promote their strategic interests. This ability to shape the international cyber environment allows states to create a favorable setting for their cyber operations and defend against potential adversaries. By promoting norms such as state sovereignty in cyberspace, the prohibition of certain types of cyber-attacks, or the protection of critical infrastructure, states can contribute to a more stable and secure international cyber landscape.

In addition, robust cyber-political capacity enables states to build and sustain strategic alliances and partnerships. These relationships can enhance a state's cyber capabilities through shared intelligence, collaborative defense initiatives, and coordinated responses to cyber threats. For example, alliances such as NATO have recognized cyberspace's significance as a warfare domain, and member states benefit from collective defense measures and shared resources to bolster their individual and collective cyber defenses.

In conclusion, while possessing state cyber capacity is integral to achieving state cyber power, realizing robust state cyber capacity requires substantial data resources and a strong digital economy. Thus, the three elements of cyber power, data resources, economic strength, and cyber political capacity are mutually reinforcing and complementary. A state's cyber-political capacity is pivotal in this triad, enabling effective utilization and governance of cyber resources to project power, protect national interests, and influence the global cyber order.

Cyber political capacity ensures that a state's cyber efforts are strategically guided, well-coordinated, and effectively implemented, thereby maximizing the potential of its cyber assets. This capacity strengthens national security and economic resilience and provides

a platform for influencing international cyber policies and norms, creating a favorable global environment for the state's cyber activities. In essence, robust cyber political capacity is the linchpin that enables states to harness their cyber resources fully, navigate the complexities of the digital age, and maintain a competitive edge in the international arena.

6. Conclusion

This study underscores the profound significance of cyberspace in contemporary global politics and the necessity of understanding cyber power within the framework of traditional IR theories. The research aims to fill a critical gap in the existing literature by applying PTT to cyberspace, considering the state as a rational and unitary actor. By integrating PTT with the rational actor model and defining cyber power based on specific objectives, this study offers a novel perspective on the assessment and categorization of cyber power.

The primary objective of this research is to define cyber power and propose a metric for its assessment, analogous to PTT's approach to evaluating national power. This involves a comprehensive analysis of cyber power by breaking it down into three core components: data resources, digital economic strength, and cyber political capacity. These elements form the basis for assessing state cyber power and understanding the hierarchical structure of states in cyberspace.

The study employs a methodological framework borrowed from PTT to achieve these objectives. It uses the rational actor model, which assumes that states act logically and strategically to maximize their interests in cyberspace. By taking the state as a unitary actor, the research simplifies the complex interactions within cyberspace, allowing for a clearer analysis of state behavior and cyber power dynamics. Furthermore, the study defines cyber power based on specific objectives, such as data ownership, information management, cybersecurity, offensive capabilities, cyberinfrastructure, and the economic aspects of cyber power.

The application of PTT to cyberspace reveals a nuanced understanding of cyber power. PTT emphasizes the importance of a comprehensive assessment of national power, traditionally measured through economic, military, and demographic indicators. This translates to a tripartite model comprising data resources, digital economic strength, and cyber political capacity in the cyber domain.

Each of these elements is crucial for a state to project power and protect its interests in cyberspace.

Data resources form the backbone of cyber power, enabling states to gather, analyze, and leverage information for strategic purposes. A strong digital economy provides the financial and technological infrastructure to support advanced cyber capabilities. However, the linchpin of this triad is cyber-political capacity. This dimension pertains to the state's ability to effectively govern and manage its cyber resources, craft coherent cyber policies, and engage in international cyber diplomacy. States with robust cyber-political capacities can coordinate their cyber activities, safeguard national security, and influence global cyber norms to create a favorable environment for their operations.

The hierarchical model of cyber power proposed in this study categorizes states into four distinct groups: global cyber leaders, cyber great powers, cyber-dependent powers, and non-cyber powers. This classification reflects the varying degrees of cyber capability and influence among states, providing a structured framework for analyzing the global cyber landscape. Global cyber leaders, or in other words, the most dominant state in cyberspace, exemplified by the United States, possess comprehensive cyber capabilities and play a central role in shaping international cyber policies. Cyber great powers, such as the European Union, China, and Russia, hold substantial influence but exhibit different levels of satisfaction with the existing cyber order, influencing their international cyber strategies. Cyber-dependent powers, while having certain cyber capabilities, rely significantly on external technologies and face vulnerabilities in cybersecurity. Non-cyber powers, with limited internet penetration and digital infrastructure, struggle to participate meaningfully in the global cyber arena.

In addition to the contribution of this study to the literature on cyberspace and IR, important clarification is also necessary regarding the scope and intent of this study. While Power Transition Theory traditionally deals with the dynamics of power shifts between states, this research does not focus on the concept of "power transition" within cyberspace. Instead, its primary aim is to apply PTT's national power model to the cyber domain to define and measure cyber power, thereby establishing a hierarchical order of states in cyberspace. This initial step is critical as it lays the groundwork for future studies to explore the dynamics of power transitions once cyber power has been accurately measured using the proposed

model. Some might see this as a limitation of the study, and critics might argue that applying PTT to cyberspace without directly exploring “transition” dynamics is premature. However, this study is a preliminary effort to introduce an IR perspective on the definition and assessment of cyber power. By borrowing PTT’s national power definition and assessment model, this research establishes the relevance of traditional IR theories to the cyber domain. This foundational work is crucial as it sets the stage for future analyses of power transitions in cyberspace, which can only be thoroughly examined once cyber power has been accurately assessed by the model introduced in this study.

In conclusion, while possessing state cyber capacity is integral to achieving state cyber power, realizing robust state cyber capacity requires substantial data resources and a strong digital economy. The three elements of cyber power, data resources, digital economic strength, and cyber political capacity, are mutually reinforcing and complementary. A state’s cyber political capacity plays a pivotal role in this triad, enabling effective utilization and governance of cyber resources to project power, protect national interests, and influence the global cyber order. By applying traditional IR theory to the domain of cyberspace and demonstrating its applicability, this research addresses a significant gap in the existing literature. It also puts forward an innovative model for assessing cyber power and provides valuable insights into the hierarchical structure of states within cyberspace. These contributions are substantial, offering a new lens through which to understand global cyber governance and geopolitical relations in this emerging and critical domain.

References

- [1] M. Carr, *US power and the Internet in international relations: The irony of the information age*. Basingstoke and New York: Palgrave Macmillan, 2016, doi: [10.1057/9781137550248](https://doi.org/10.1057/9781137550248).
- [2] A.H. Morrison, “An impossible future: John Perry Barlow’s ‘Declaration of the Independence of Cyberspace,’” *New Media and Society*, vol. 11, no. 1–2, pp. 53–71, 2009, doi: [10.1177/1461444808100161](https://doi.org/10.1177/1461444808100161).
- [3] R. Creemers, “Governing cyberspace: behavior, power and diplomacy,” in *Governing Cyberspace: Behavior, Power and Diplomacy*, D. Broeders, B. Van den Berg, Eds., London: Rowman & Littlefield, 2020, pp. 116–151.
- [4] Z. Hongren, “Strategic stability in cyberspace: a chinese view,” *China Quarterly of International Strategic Studies*, vol. 5, no. 1, pp. 81–95, 2019, doi: [10.1142/S2377740019500088](https://doi.org/10.1142/S2377740019500088).

- [5] J. Nocetti, "Contest and conquest: Russia and global internet governance," *International Affairs*, vol. 91, no. 1, pp. 111–130, 2015, doi: [10.1111/1468-2346.12189](https://doi.org/10.1111/1468-2346.12189).
- [6] H. Ebert, T. Maurer, "Contested cyberspace and rising powers," *Third World Quarterly*, vol. 34, no. 6, pp. 1–22, 2013, doi: [10.1080/01436597.2013.802502](https://doi.org/10.1080/01436597.2013.802502).
- [7] J. van Haaster, "Assessing cyber power," in *2016 8th International Conference on Cyber Conflict*, M. V. N. Pissanidis, H. Rõigas, Eds., Tallinn: NATO CCD COE Publications, 2016, pp. 85–90.
- [8] N. Inkster, "Measuring military cyber power," *Survival*, vol. 59, no. 4, pp. 27–34, 2017, doi: [10.1080/00396338.2017.1349770](https://doi.org/10.1080/00396338.2017.1349770).
- [9] M. Willett, "Assessing cyber power," *Survival*, vol. 61, no. 1, pp. 85–90, 2019, doi: [10.1080/00396338.2019.1569895](https://doi.org/10.1080/00396338.2019.1569895).
- [10] D.J. Betz, T. Stevens, "Power and cyberspace," *Adelphi Series*, vol. 51, no. 424, pp. 35–54, 2011, doi: [10.1080/19445571.2011.636954](https://doi.org/10.1080/19445571.2011.636954).
- [11] J.S. Nye, "Cyber Power," *Harvard Kennedy School Belfer Center for Science and International Affairs*, no. 6, pp. 2–4, 2015, doi: [10.12816/0022579](https://doi.org/10.12816/0022579).
- [12] G. Rose, "Neoclassical Realism and theories of foreign policy," "<https://www.cambridge.org/core/journals/world-politics>" *World Politics*, vol. 51, no. 1, pp. 144–172, 2010.
- [13] T. Dunne, M. Kurki, S. Smith, Eds., *International Relations Theories Discipline and Diversity*, 1st ed. London: Oxford University Press, 2007.
- [14] J.S. Nye, *Soft Power and Great-Power Competition: Shifting Sands in the Balance of Power Between the United States and China*. MA Singapore: Springer. doi: <https://doi.org/10.1007/978-981-99-0714-4>.
- [15] E.L. Armistead, "Suggestions to measure cyber power and proposed metrics for cyber warfare operations (cyber deterrence/cyber power)," *2016 IEEE International Conference on Cyber Conflict, CyCon U.S.* 2017, doi: [10.1109/CYCONUS.2016.7836610](https://doi.org/10.1109/CYCONUS.2016.7836610).
- [16] J. Arquilla, D. Ronfeldt, "Cyberwar is coming," *Rand Corporation*, 1993. Available: <https://www.rand.org/pubs/reprints/RP223.html> [Accessed: July 25, 2024].
- [17] M.C. Libicki, "Information war, information peace," *Journal of International Affairs*, vol. 51, no. 2, pp. 411–428, 1998.
- [18] M. Castells, *The Rise of the Network Society*. New York: Blackwell Pub, 1996. doi: [10.1002/9781444319514](https://doi.org/10.1002/9781444319514).
- [19] S. Sassen, *Globalization and Its Discontents Essays on the New Mobility of People and Money*. New York: The New Press, 1999.
- [20] R.J. Keohane, J. Nye, *Power and Interdependence*. 1st ed. New York: Pearson, 1988.
- [21] J.N. Rosenau, J.P. Singh, Eds., *Information Technologies and Global Politics The Changing Scope of Power and Governance*. New York: State University of New York Press, 2002.
- [22] H. Nissenbaum, "Where computer security meets national security," *Ethics and Information Technology*, vol. 7, no. 1, pp. 61–73, 2005, doi: [10.1007/s10676-005-4582-3](https://doi.org/10.1007/s10676-005-4582-3).

- [23] A. Chadwick, *Internet Politics: States, Citizens, and New Communication Technologies*, 1st ed. London: Oxford University Press, 2006.
- [24] R. Deibert, R. Rohozinski, "Control and subversion in Russian cyberspace," in *Access Controlled: The Shaping of Power, Rights, and Rule in Cyberspace*, R. Deibert, J. Palfrey, R. Rohozinski, J. L. Zittrain, Eds., London: The MIT Press, 2010, pp. 15–34. doi: <https://doi.org/10.7551/mitpress/8551.001.0001>.
- [25] N. Choucri, *Cyberpolitics in International Relations*. London: The MIT Press, 2012. doi: [10.7551/mitpress/7736.003.0009](https://doi.org/10.7551/mitpress/7736.003.0009).
- [26] J.R. Lindsay, "Stuxnet and the limits of cyber warfare," *Security Studies*, vol. 22, no. 3, pp. 365–404, Jul. 2013, doi: [10.1080/09636412.2013.816122](https://doi.org/10.1080/09636412.2013.816122).
- [27] A. Venables, S.A. Shaikh, J. Shuttleworth, "The projection and measurement of cyberpower," *Security Journal*, vol. 30, no. 3, pp. 1000–1011, 2017, doi: [10.1057/sj.2015.35](https://doi.org/10.1057/sj.2015.35).
- [28] D. Brizhinev, N. Ryan, R. Bradbury, "Modelling hegemonic power transition in cyberspace," *Complexity*, vol. 2018, pp. 1–13, 2018, doi: [10.1155/2018/9306128](https://doi.org/10.1155/2018/9306128).
- [29] R.L. Tammen, *Power transition: Strategies for the 21st century*, 1st ed. New York: CQ Press, 2000.
- [30] W. Kim, S. Gates, "Power transition theory and the rise of China," *International Area Studies Review*, vol. 18, no. 3, pp. 219–226, 2015, doi: [10.1177/2233865915598545](https://doi.org/10.1177/2233865915598545).
- [31] M. Bussmann, J. R. Oneal, "Do hegemons distribute private goods?: A test of power-transition theory," *Journal of Conflict Resolution*, vol. 51, no. 1, pp. 88–111, 2007, doi: [10.1177/0022002706296178](https://doi.org/10.1177/0022002706296178).
- [32] R.N. Lebow, B. Valentino, "Lost in transition: a critical analysis of power transition theory," *International Relations*, vol. 23, no. 3, pp. 389–410, 2009, doi: [10.1177/0047117809340481](https://doi.org/10.1177/0047117809340481).
- [33] R. L. Tammen, J. Kugler, D. Lemke. (Oct. 26, 2017). Foundations of Power Transition Theory. [Online]. Available: <https://oxfordre.com/politics/display/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-296?product=orepol>. [Accessed: Apr. 04, 2021].
- [34] W. Kim, S. Gates, "Power transition theory and the rise of China," *International Area Studies Review*, vol. 18, no. 3, pp. 219–226, 2015, doi: [10.1177/2233865915598545](https://doi.org/10.1177/2233865915598545).
- [35] Y. Akdag, "The likelihood of cyberwar between the United States and China: A neorealism and power transition theory perspective," *Journal of Chinese Political Science*, vol. 24, no. 2, pp. 225–247, 2019, doi: [10.1007/s11366-018-9565-4](https://doi.org/10.1007/s11366-018-9565-4).
- [36] M. Bey, "Great powers in cyberspace: the strategic drivers behind US, Chinese and Russian competition," *International Conference on Cyber Conflict*, vol. 126, no. 1, pp. 1–7, 2019.
- [37] T. Ray, "The quest for cyber sovereignty is dark and full of terrors," *ORF*. [Online]. Available: <https://www.orfonline.org/expert-speak/the-quest-for-cyber-sovereignty-is-dark-and-full-of-terrors-66676/>. [Accessed: Apr. 04, 2021].
- [38] T. Woodhouse, "Alliance for affordable internet," 1BC. [Online]. Available: <https://a4ai.org/report/2021-affordability-report/#i-acknowledgements.2021>. [Accessed: Dec. 23, 2023].

- [39] S. Han, "China's pursuit of peaceful power transition: a case of ICT (Information and Communications Technologies) standard setting," *International Area Studies Review*, vol. 12, no. 3, pp. 27–42, 2009, doi: [10.1177/223386590901200302](https://doi.org/10.1177/223386590901200302).
- [40] A. Kosenkov, "Cyber conflicts as a new global threat," *Future Internet*, vol. 8, no. 3, 2016, doi: [10.3390/fi8030045](https://doi.org/10.3390/fi8030045).
- [41] R.B. Andres, "The Emerging Structure of Strategic Cyber Offense, Cyber Defense, and Cyber Deterrence," in *Cyberspace and National Security: Threats, Opportunities, and Power in a Virtual World*, D. S. Reveron, P. Jagoda, H. Lin, Eds., Washington DC: Georgetown University Press, 2012, pp. 89–104.
- [42] 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](https://doi.org/10.1162/ISEC_a_00136).
- [43] J. Zhang *et al.*, "The impact of digital economy on the economic growth and the development strategies in the post-COVID-19 era: evidence from countries along the 'Belt and Road,'" *Frontiers in Public Health*, vol. 10, no. May, pp. 1–17, 2022, doi: [10.3389/fpubh.2022.856142](https://doi.org/10.3389/fpubh.2022.856142).
- [44] M. Carr, "Power plays in global internet governance," *Millennium: Journal of International Studies*, vol. 43, no. 2, pp. 640–659, 2015, doi: [10.1177/0305829814562655](https://doi.org/10.1177/0305829814562655).
- [45] J.J. van Vuuren, L. Leenen, "A model for measuring perceived cyberpower," Proceedings of the 13th International Conference on Cyber Warfare and Security, ICCWS 2018, 2018, pp. 320–327.
- [46] N. Inkster, *China's Cyber Power*, 1st ed. Oxon: Taylor & Francis, 2016. Available at: <https://www.routledge.com/Chinas-Cyber-Power/Inkster/p/book/9781138211162> (Accessed: Jan. 20, 2023).
- [47] L. Tsui, "The panopticon as the antithesis of a space of freedom: Control and Regulation of the Internet in China," *China Information*, vol. 17, no. 2, pp. 65–82, 2003, doi: [10.1177/0920203X0301700203](https://doi.org/10.1177/0920203X0301700203).
- [48] R.Á. Pinto, "Digital sovereignty or digital colonialism?," *International Journal on Human Rights*, vol. 15, no. 27, pp. 15–27, 2018.
- [49] S. Hoffmann, S. Bradshaw, E. Taylor, "Networks and geopolitics: How great power rivalries infected 5G," *Oxford Information Labs*, pp. 37, 2019.
- [50] F.C. Domingo, "Conquering a new domain: Explaining great power competition in cyberspace," *Comparative Strategy*, vol. 35, no. 2, pp. 154–168, 2016, doi: [10.1080/01495933.2016.1176467](https://doi.org/10.1080/01495933.2016.1176467).
- [51] T. Rid, B. Buchanan, "Attributing cyber attacks," *Journal of Strategic Studies*, vol. 38, no. 1–2, pp. 4–37, 2015, doi: [10.1080/01402390.2014.977382](https://doi.org/10.1080/01402390.2014.977382).
- [52] S.W. Loneragan, "Cyber power and the international system," Columbia University, 2017. doi: [10.7916/D88D07PH](https://doi.org/10.7916/D88D07PH).
- [53] UNCTAD, "Digital Economy Report 2021- Cross-border data flows and development: for whom the data flow," New York, 2021.

- [54] Q. Meng, M. Li, "New economy and ICT development in China," *Information Economics and Policy*, vol. 14, no. 2, pp. 275–295, 2002, doi: [10.1016/S0167-6245\(01\)00070-1](https://doi.org/10.1016/S0167-6245(01)00070-1).
- [55] Centre for Strategic and International Studies, *G20 Toolkit for Measuring Digital Skills and Digital Literacy: Framework and Approach*. [Online]. Available: <https://www.csis.or.id/publication/g20-toolkit-for-measuring-digital-skills-and-digital-literacy-framework-and-approach/> (Accessed Apr. 02, 2023).

Ransomware: Why It's Growing and How to Curb Its Growth

Joshua Jaffe | Oxford Internet Institute, Oxford University, UK |
ORCID: 0000-0001-8238-4949

Luciano Floridi | Digital Ethics Center, Yale University, USA |
ORCID: 0000-0002-5444-2280

Abstract

Ransomware is an increasingly pernicious threat to individuals, businesses, economies, and societies. Ransomware attacks simplify the typical cybercrime value chain. Given the exponential growth of data, the wide distribution of connected devices, the so-called internet of things, and the power of artificial intelligence to exponentially scale attacks, ransomware is likely to continue to grow. Much research and analysis has focused on ransomware tool kits, malware samples, and the vulnerable victim landscape. However, this is only part of the picture. At its core, ransomware is a crime committed almost entirely for economic benefit. Yet, research on behavioural factors and market forces that incentivise the proliferation of ransomware is limited. The majority of what does exist comes in the form of media reporting and industry periodicals. Given their relevance, these sources should not be discounted out of hand. Yet, how critically should their findings be viewed and inherent conflicts within their findings be resolved? Further, as the profit motive of ransomware is similar to other economic crimes, how relevant is the vast body of research on criminality or on behavioural economics to understanding the growth of ransomware? In this article, we review the literature relevant to understanding the growth of ransomware by widening the lens to include a range of relevant multi-disciplinary academic sources as well as industry data. We then discuss our conclusions regarding

Received: 11.06.2024

Accepted: 07.08.2024

Published: 09.11.2024

Cite this article as:

J. Jaffe, L. Floridi
"Ransomware: Why it's growing and how to curb its growth," ACIG, vol. 3, no. 2, 2024, pp. 72–98. DOI: 10.60097/ACIG/192959

Corresponding author:

Joshua Jaffe, Oxford
Internet Institute, Oxford
University, UK; E-mail:
joshua.jaffe@mansfield.
ox.ac.uk

 0000-0001-8238-4949

Copyright:

Some rights reserved

(CC-BY):

Joshua Jaffe
Luciano Floridi
Publisher NASK



the forces compelling its growth and identify areas requiring further study that could reverse the trend.

Keywords

Ransomware, cybercrime, cyber warfare, extortion, malware

1. Introduction

The ransomware trend in cybercrime is growing. Online virus database VirusTotal has received uploads of more than 80 million ransomware samples since 2020 [1]. According to global telecommunications company Verizon, the frequency of ransomware attacks doubled in 2021 [2]. In its survey from the same year, the International Data Corporation (IDC) found that 37% of companies reported having been the victim of ransomware, the highest percentage in the survey's history [3]. The Federal Bureau of Investigation reports that ransomware-related complaints have risen 62% year-on-year in the United States [4]. The World Economic Forum considers cybercrime the most significant threat to businesses in the United States, Canada, and Europe [5]. In 2020, Farahbod et al. estimated the cost of cybercrime on the global economy at 'up to \$1 trillion' [6]. Editor-in-chief of *Cybercrime Magazine* Steve Morgan went further, estimating the overall cost of cybercrime would exceed \$10 trillion by the end of 2025. He also noted that ransomware is increasingly becoming the go-to choice for cybercriminals [7].

Most of the data about the cost of cybercrime and the growth of ransomware come from industry sources. Though the above statistics are staggering in their claims, it should be noted from the outset that the methodology for calculating the cost of cybercrime, or a particular variety like ransomware, varies considerably by author. Further, many of these industry sources have a vested interest in certain perceptions of ransomware crime, so – while they fill a gap in the literature – their findings should be subject to skepticism. Anderson and coauthors address some of these challenges in their 2019 reprisal of their 2012 paper, noting that, in addition to challenges with availability of data, there is also a methodological issue as well [8]. They note some authors include only the direct losses to hackers, others consider the indirect societal costs and the invisible tax passed along to consumers in the form of growing cybersecurity budgets that inevitably find their way to the cost of goods sold [8].

Regardless, without attempting resolve the precise societal cost of ransomware, the growth of this crime observed by all the above

sources suggests that ransomware has become endemic, and this has far-reaching implications for individuals, corporations, societies, and economies. Despite the alarming increase in ransomware, the underlying social and ethical forces involved remain understudied. Most analyses focus on *what* and *when* questions. They enumerate the details of isolated attacks and adopt a technical approach to analysing specific malware tool kits and individual criminal actors. This kind of research is necessary but insufficient because it grants only a partial understanding of the growing ransomware phenomenon. It is inadequate for drawing societal conclusions to address the problem because it does not consider the actors' *motivations* and *values*.

Conventional approaches limit our capability to circumscribe and minimise ransomware attacks because they provide an incomplete understanding of the scope and scale of the problem. This is inconsistent with how we usually address other social and criminal ills. Typically, policymakers focus on *who*, *why*, and *how* questions. For example, law enforcement does not develop strategies for combating violent crime by evaluating individual shootings and context-specific forensic evidence from an individual event. Public safety officials do not write building codes based on a detailed study of an individual residential fire. Nor do national security officials develop strategy solely based on an individual adversary's infantry forces. Stated this way, common sense, and general familiarity with each broad category of policy, make the above examples unsuitable for drawing macro conclusions about combatting violent crime, improving residential building standards, or securing a national defence. Each of these domains is composed of a mosaic of factors, and the relevant actors have a complex range of motivations. Effective policing strategy considers the motivations of criminal actors and the forensic specifics of individual crimes. Fire prevention requires the thoughtful selection of materials, construction in accordance with building code requirements, and responsible behaviour on the part of individuals. Defence policy does not rely solely on analyses of an adversary's military capabilities but also on national interests and the character of their respective leaders. As a result, in this review, we widen the aperture and consider a range of literature relevant to better understanding the motivations of ransomware actors as well as the scale of ransomware crime.

1.1. Scope of Analysis

Individuals, governments, and societies solve systemic problems by understanding and addressing all the relevant factors

that drive behaviour. Also visible against the backdrop of these examples is the fact that the success of the policy is not dependent on criminal code and legal redress alone. They also depend on a degree of convergence between the norms and values within a society and the problem in question. For example, most individuals in a society do not merely avoid criminal behaviour because it has been defined as illegal but also because social pressures are applied which cause a criminal record to carry a social penalty. Construction companies don't comply with civil building codes simply because they are legally required to but also because there are commercial penalties associated with a poor safety reputation. And, a strong national defence is not merely the product of defensive arms but also strong alliances, social cohesion, and economic resilience. When all works well, this can align individual motivations with desirable ends, such as social progress and the collective good.

This article is intended to provide the grounds for future analyses of how the growth of ransomware might be curtailed through socio-economic interventions. In doing so, we aim to (1) provide a systematic overview of the problem, (2) assess the state of the current debate, and (3) suggest underexplored areas of both practical and theoretical interest for tackling the ransomware problem. We focus on governance, ethical, legal, and social implications (GELSI). We also engage with well-studied cases from the social sciences relevant to our topic (e.g. issues around paying conventional ransoms to kidnappers).

In our analysis, we focus on (1) single ransomware, which refers to the encryption of data and then the holding of the decryption keys for ransom, and (2) double ransomware, which is like single ransomware but with the addition of extortion involving the public disclosure of stolen data to compel ransom payments [9]. These two types of ransomware account for most of its growth. They also share a common motivation: compelling a data owner or custodian to act against their interests through extortion. We do not discuss purely destructive cyberattacks, nor do we discuss so-called false-flag ransomware, which disguise attacks intended to be purely destructive as ransomware attacks [10]. We make this distinction because we consider destructive cyberattacks and false-flag ransomware to be different kinds of phenomena because the motivations of the actors are different.

1.2. Methodology

Motivated by an interest in understanding the forces driving ransomware's growth, we conducted a *state of the art review* of

the literature relevant to the social and behavioural analysis of ransomware crimes [11]. We structured our review of the relevant literature by focusing on articles that can help answer the *who*, *why*, and *how* questions. We reviewed more than 100 sources and ultimately selected 50 for inclusion on the basis of their novelty and relevance to answering these questions. We relied on academic journal articles that describe the origin and nature of ransomware crimes committed over the course of the past four decades. However, this review also subjects a wide range of industry research and statistics on ransomware to critical review. While we did ultimately include some industry estimates of the scope and scale of ransomware we considered most credible, we focused primarily on those sources able to help characterise the behaviour of cybercriminals and answer the *who*, *why*, and *how* questions noted above.

Given that ransomware has many similarities with conventional economically motivated crimes, this review also considers literature in the fields of Criminology and Economics that we believe adds to the collective understanding of ransomware's growth. We conducted further analysis, applying conventional techniques used in these disciplines to reach indirect conclusions about these ransomware questions, where no direct contextual data relevant to a specific aspect of the ransomware problem was uncovered through our research. Finally, we also interviewed some experts, including their insights into our findings (see Figure 1).

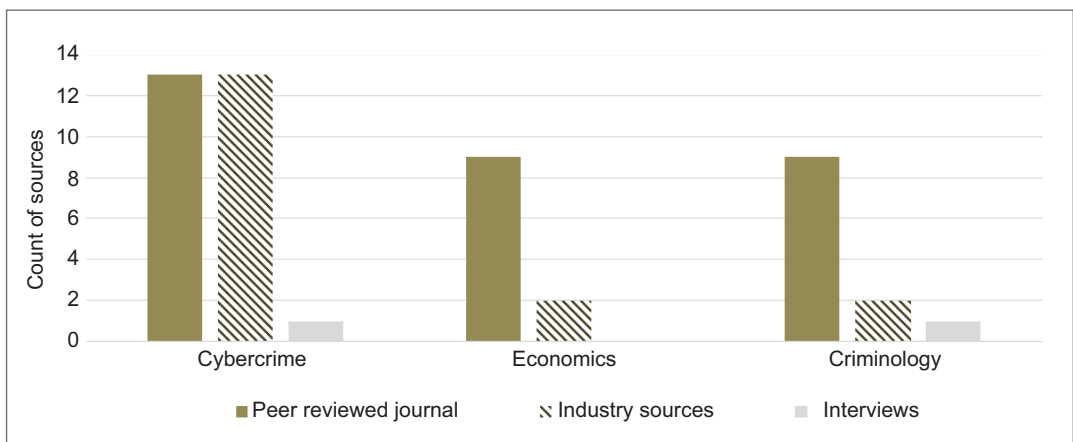


Figure 1. Included sources by field of study.

1.3. Structure and Framing of Analysis

Our findings organise the literature on the social forces involved in the rise of ransomware into five sections. Following the

introduction found in Section 1, we proceed to Section 2, where we describe the *what* and *when* questions about ransomware. This establishes the foundational claim that ransomware is an endemic problem. In Section 3, we address the *who* question by reviewing the research on ransomware actors and the origin of attacks. This supports the claim that actor-analyses are mostly descriptive and lack an understanding of motivating factors. In Section 4, we focus on the *why* question. We consider the literature on the illicit marketplace for ransomware, the exchange of value in the marketplace, and how this drives actor behaviour. In Section 5, we discuss the *how* question. We evaluate the effects of various practices on the ransomware problem and the adverse selection bias involved. In Section 6, we conclude our analysis and consider some areas for further study.

2. What and When: A Brief Summary of the Evolution of Ransomware Tactics

The increase in ransomware attacks may have surprised many in government and industry, but the core reason for such growth is not a mystery. Ransomware attacks simplify the typical cybercrime value chain, where reduction in complexity drives growth. In this section, we discuss how ransomware has been employed as an attack method to extract value over the past two decades.

Ransomware attacks were common but not epidemic until 2013. Since 2013, they have grown by more than 500% [12]. Ransomware evolved in the late 1990s from simple user interface (UI)-lockers to disk-encrypting cryptographic ransomware. More recently, they have advanced to include file-exporting tool kits that encrypt users' data and enable data theft [13]. For over a decade, most attacks opportunistically targeted individuals, typically with random mass-mail Spam or indiscriminate drive-by downloads.¹ Over this period, almost all attacks originated in Russia, and targets were mostly in Russia or countries on the Russian periphery [14].

The number of ransomware malware samples doubled each quarter in 2011, mainly owing to the development of commercialised ransomware tool kits and anonymous payment systems [14]. This sharp growth continued as the illicit market for ransomware tool kits, know-how, and payment mechanisms expanded. Ransomware attacks exploded in 2016 when there was a tactical shift towards targeting large corporations with so-called wormable ransomware (ransomware that can burrow through a computer network without direct control from the hacker). This naturally correlated with

1——A drive-by download is a method of exploiting a victim computer that can infect a vulnerable web browser software if a user visits a compromised website.

increasingly high ransom demands. This coincided with a rise in cyber extortion over the same period of time, where not only was data held for ransom but the threat of exposing to regulators the fact that an organisation had been hacked is used to compel speedy payment [15].

Only a negligible fraction of this reported growth can be explained by improved methods for detecting ransomware attacks. The online computer virus aggregator VirusTotal counts 11.7 billion ransomware malware samples uploaded to its services since 2005 [1]. When plotted over time, the increase represents a growing wave, rather than a sudden jump. Leveraging data collected and reported by Verizon, we find that ransomware accounted for less than 1% of all reported cyberattacks in 2013 but more than 25% in 2021 (see Figure 2) [16]. The compounded annual growth rate (CAGR) for this period exceeds 50% per annum, a significant increase and one that supports the observed trends.

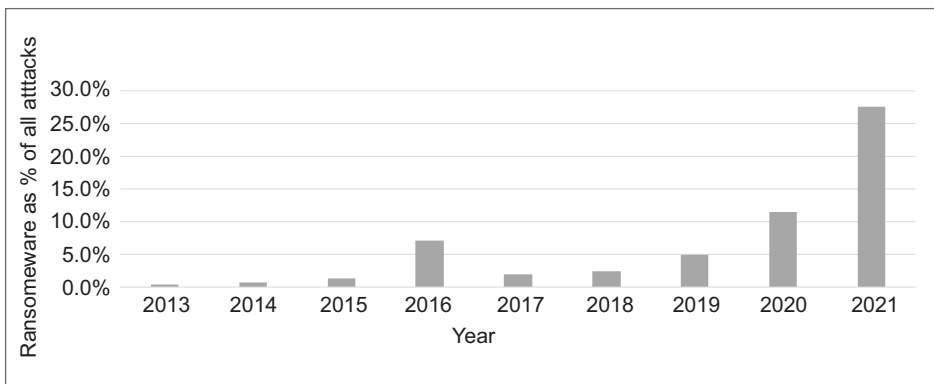


Figure 2. Ransomware as Percent of Total Reported Attacks in Verizon VERIS Data Base and DBIR Report.

That said, a closer look at the data suggests a couple of caveats. Firstly, although the general trend in reported ransomware crimes has trended up, the overall number of cybercrimes reported in VERIS has declined consistently since 2013. Part of the ransomware's annual percentage growth could be attributed to this decline in the denominator.² Secondly, significant regulation in this period created new reporting and remuneration obligations for corporations affected by ransomware. This likely impacted the number willing to publicly report any cyberattack, possibly resulting in gross undercounting. Although it is impossible to know for sure, we think that the decline in general cybercrimes being reported

²—At the time of these analyses, the raw total of attacks in VERIS was available only for the year 2017. The 2022 DBIR Report, which is calculated on the raw VERIS data, provided the percentage of attacks categorised as ransomware from 2017 to 2021. It is therefore possible to complete the table in Figure 1 only as a comparison of percentages.

is attributable to a shift in emphasis away from nuisance crimes towards more significant cases. Given several reporting disincentives, it seems likely that ransomware events are undercounted. We suspect that the trend in ransomware crimes is more severe than represented in Figure 2.

Regardless of which sources are consulted, ransomware is a prevalent and growing method by which cybercriminals seek to extract value. It also appears that the wave has not yet crested. In their survey of ransomware techniques, McIntosh and colleagues note that there is a consensus expectation that ransomware attacks will not only continue to grow but also shift towards more disruptive tactics that are more difficult to combat [13]. According to McIntosh et al. [13],

1. There will be a reduction in attacks on private individuals and an increase in attacks on organisations, further optimising the time-to-value ratio in favour of the attackers.
2. There will be a shift in tactics towards active exploitation of technology vulnerabilities and away from passive infiltrations (e.g. via phishing, vishing, or fraud).
3. There will be a broadening of the mechanism to deprive enterprises of access to their systems, possibly renewing the focus on distributed denial of services (DDoS) attacks instead of only file encryption.

With these forecasts as the backdrop, the proliferation of network-connected industrial internet of things (IIoT) devices upon which vital social enterprises rely raises stark concerns. Many of these have been summarised by Yaqoob et al. [17], who stress the vital functions that connected IIoT devices perform. These devices have also proven to be significantly vulnerable to ransomware attacks. At the macro level, Yaqoob et al. discuss ransomware risks to hospital centres, water treatment facilities, the electrical grid, pharmaceutical production, and nuclear reactors [17]. At the micro level, autonomous vehicles and implantable medical devices appear particularly at risk. Society is becoming increasingly dependent on technology, and connected devices play an increasingly vital role in human safety and societal well-being. It is insufficient to consider ransomware attacks within the limited view of technical exploits and countermeasures. Ransomware requires a response similar to approaches addressing other grave societal threats. Such a response, we contend, must recognise the motivations of the bad actors involved and realign their interests with those of society at large.

3. Who: On Actors, Motivations, and Public Perception

Data on cybercriminals is difficult to gather, given the shadowy and opaque nature of cybercrime. As such, cybercriminals' motivations can be challenging to assess and categorise. These issues make it difficult for the public to perceive the problem accurately. It also makes developing a framework or standard for ethical behaviour challenging. Cyberattacks also have effects beyond strictly financial ones. For example, ransomware attacks against hospitals in the United States and Europe have prolonged patients' wait times for critical care.³ Nonetheless, public perception and ethical criticism of cybercriminals and cybercriminal activity remain mixed owing to the cyber domain's opacity. Social pressure on cyber criminals also remains only mildly influential.

3——In one case, a patient in a German hospital died while waiting for emergency treatment [5].

3.1. Public Perception of Cybercrime and Ransomware

Mulhall's survey of public perceptions of cybercriminals is dated, but it shows some interesting trends [18]. When viewed from the largely benign perspective of hacking, public perceptions tend to be mixed. Many survey respondents had a negative association with terms connected with cybercriminal behaviour. This was most closely associated with news of attacks that personally affected people. Negative associations were especially acute when attacks risked the health or lives of individuals. However, when hackers targeted nameless/faceless corporations, especially those with poor public reputations, then public opinion was less condemning (Mulhall focuses on the targeting of the US and British Telcom giants at the height of their profits).

Given the age of this survey, we should supplement it with more recent corroboration. There is evidence to suggest a parallel in current public sentiment. Pawlicka and colleagues illustrate this by citing examples of so-called hacktivism [19]. Hacktivism targets organisations that the hackers believe are perpetrating a systemic injustice. As such, most attacks do not cause general public alarm. Harford, from marketing and sales services company TechTarget, notes that, prior to 2016, ransomware attacks were mostly limited in scope and sophistication [20]. They targeted individuals, ransom-ing personal files, photos, and financial documents. Attackers often adopted a friendly approach, sometimes even apologising for the inconvenience and offering support to fix the problem after the ransom was paid [20]. There is not much literature on the effect of this tactic, but public outrage was generally muted. This changed in 2016 with the *Petya* and *WannaCry* attacks. These attacks leveraged

the EternalBlue exploit, which allows malware to worm through a victim's network. This new capability led to the targeting of enterprises, large-scale damage, and the extraction of much larger ransoms [4]. This and other similar tactics increased the scope of the attacks and led to wider ripples throughout the society, often affecting individuals well beyond the targeted company. Examples include the attack on the British NHS in 2017, staple food producer JBS in 2021, and the energy company Colonial Pipeline in 2021. These directly impacted consumers' convenience, health and/or financial well-being. The result was stark shift in perceptions of this kind of crime and the actors who perpetrate it [21].

Applying these findings to the modern ransomware context leads to two conclusions, both suggesting the need for further study. Firstly, public outrage was limited when wealthy corporations were targeted and where members of the public were not directly impacted (either financially or socially) [18]. Secondly, this sentiment reverses after 2015. This correlates with a shift to more risky tactics, more impactful and prominent targets, and increased public concern.

3.2. Motivations of Cybercriminals

Direct, first-person accountings of what motivates those involved in ransomware or other types of cybercrime often suffer from bias. Journalistic reporting about those engaged in this type of criminal activity is often overly influenced by a few sensational cases. They range from the comical Kindergarten Hacker [22] to the legendary Evil Corp [23]. However, a few more grounded analyses do exist, which provide some insights about motivations.

A 2016 analysis of self-described hackers from the United States, the United Kingdom, and Germany was conducted by PaloAlto Networks and the Ponemon Institute. They found that most cybercriminals fit the stereotype. Most were underemployed (the average annual income from cybercrime was slightly more than £20,000). More than two-thirds claimed that monetary gain was their sole or primary motivation. On average, they completed only two successful attacks per year. These were, however, sufficiently lucrative to make the attacks worth the investment of time and resources. The typical attack took less than 24 hours to execute and yielded an average return of between £8,600 and £10,900, depending on the country of the respondent [24].

Security periodical *CSO Online* estimates that the aggregate cost of cybercrime likely exceeded \$6 trillion in 2021 [25]. Similar surveys

also provide some insights into cybercriminal motivations. However, they likely suffer from sampling error. For both the Ponemon Institute and the *CSO Online* estimates to be correct, approximately one in ten people would need to be engaged in cybercrime. This seems highly implausible. Ian Thornton-Trump [26] offers a more likely explanation. Many cybercriminals are freelancers, but most losses result from professional cybercriminals working full-time. Professional cybercriminals use much more sophisticated methods and therefore cause much more damage. Most are organised into criminal cartels [16]. They share the profit motive identified by PaloAlto Networks and the Ponemon Institute but execute their attacks more frequently and precisely [24]. Further, Gragido et al. shed some light on the big business of cybercrime. They demonstrate the approach of mature syndicates taking a structured approach to cybercrime research and development (R&D), often investing millions of dollars with the realistic prospect of achieving many millions more in return on their investments [27].

4. Why: On the Marketplace for Ransomware

Cybercrime Magazine calculated that the cost of ransomware grew from \$325 million in 2015 to \$5 billion in 2017 [28], an increase of more than 1500%. According to the threat research team at Verizon, ransomware attacks represented 3% of all cyberattacks in 2017 [2]. By the end of 2021, ransomware attacks accounted for 25% of all cyberattacks. The associated value lost is estimated to grow to an aggregate of \$265 billion by the end of the decade [29]. This, too, likely represents a significant underestimation of the damages due to the severe disincentives to public reporting of ransomware attacks noted above.

Interestingly, the illicit trade in ransomware malware seems quite efficient despite the large volume of malicious ransomware code. Cyber actors, like conventional actors, engage in a rational evaluation of tradeoffs before choosing to commit a crime. This is consistent with application of the Rational Choice Theory, now widely applied to other conventional crimes [30]. Ransomware exhibits higher benefits and lower costs than other types of cybercrime. The macro factors driving the growth of ransomware (apart from other types of cybercrime) appear to be related to its ability to convert criminal activity into value efficiently. Historically, cybercriminals needed to go through the following nine steps: (1) discover a vulnerability in a system, (2) create malware capable of exploiting the vulnerability, (3) 'weaponised' that malware to gain access to a victim system, (4) conduct 'reconnaissance' until data considered valuable is recognised, (5) exfiltrate

those data without being blocked, (6) market the data for sale at illicit marketplaces, (7) find a prospective buyer, (8) gain the buyer's trust regarding validity and uniqueness of the data, and finally (9) conduct an exchange of value. Contrast this with ransomware, where the data can be assumed to be valuable because they are currently being used by the custodian, nothing needs to be exfiltrated, and the buyer is built into the equation from day one.

4.1. The Market Concentration of Ransomware Malware

This efficiency does not stop with the attack itself; it extends into the ransomware 'ecosystem'. Analysing the data reported by VirusTotal, it appears that the commercial hacker market operates in a near-frictionless, highly consolidated fashion, where capital is allocated to the most efficient software. Traditionally, economists use the Herfindahl-Hirschman Index (HHI) to assess market concentration. The HHI sums the square of each vendor's market share in a market segment. It does so by using the following simple formula: $HHI = s_1^2 + s_2^2 + \dots + s_n^2$, where s denotes market share and n denotes the number of competitors in the market. When evaluating monopolistic market power in anti-trust cases, the US Department of Justice considers an HHI of more than 2500 to be highly concentrated. If we apply the HHI model to the selection of ransomware malware samples reported by VirusTotal, then we get an HHI score of 6250 (see Table 1). This

Table 1. Top ransomware families as percentage of total reported ransomware malware samples described to VirusTotal.

Top 10 malware families	% of Samples	HHI score
Gandcrab	78.5%	6162.3
Babuk	7.6%	57.9
Cerber	3.1%	9.7
Matsnu	2.6%	6.9
Wannacry	2.4%	5.8
Congur	1.5%	2.3
Locky	1.3%	1.7
Teslacrypt	1.1%	1.3
Rkor	1.1%	1.2
Reveton	0.7%	0.5
Total	100.00%	6249.5

is more than 2½ times the Department of Justice bar for highly concentrated. Of more than 60 million samples organised into 130 malware families in 2020, cybercriminals chose the *Grandcrab* malware more than 75% of the time. The top three malware families accounted for approximately 90% of all attacks; malware families 11–130 accounted for less than 1% of all attacks.

The frequency with which cybercriminals use a piece of malware is only partially attributable to functionality and vulnerabilities exploited. A tool kit's flexibility for payment mechanisms and the built-in ability to obscure traceability are also important. Kharraz and colleagues thoroughly analysed the most popular ransomware software [12]. They reached some interesting conclusions about attacker behaviour. Analysing 1359 samples, they found that more than 80% of tool kits included features for obscuring payment traceability. Not surprisingly, cryptocurrencies were most popular for receiving extorted money, with bitcoin being the cryptocurrency most demanded by attackers at the time of the study. Others requested cash cards, like Moneypak, Paysafe, or UKash. Of those using bitcoin, almost three-quarters used a bitcoin address for only two transactions (the incoming transaction to receive payment, then an outgoing one to move the funds) [12]. From there, attackers split the outgoing funds into multiple accounts (or cryptocurrency wallets) to obscure traceability. They laundered the extorted funds by mixing them with funds in other wallets accumulated from various sources. The 'clean' funds were later recombined and dispersed back to the attacker in a 'clean wallet'. Most of the accounts and aliases associated with these wallets were active for fewer than five days. Following this period of time, they were often discarded and never used again.

4.2. Component Costs and Value Creation of Ransomware Tools

The darkweb marketplace for the different components of a ransomware attack is opaque but not impossible to survey. Huang and colleagues offer clues on how value can be exchanged and disrupted. They document entire pharmacy databases of customers' personal information available for less than \$1000 [31]. There are groups (or so-called bot-nets) of compromised devices with pre-installed bitcoin mining software for an average price of €2.25. Phishing services, managed by professional cybercriminals and operating on a criminal customer's behalf, cost approximately \$100 per month [31].

Although their research into the value chain of ransomware transactions was limited to a single example, it provides some evidence that warrants broader study. Huang et al. note that the darkweb purchase of the Neutreno ransomware payload, corresponding tool kit, and related services to execute a ransomware attack end-to-end would cost approximately \$13,000 per/month plus an aggregate commission of 40% on gains. Based on reports by the Cisco cyber research team, conservative estimates of return on investment by a skilled hacker gang would exceed 500% or \$81,000 per/month [31]. This could be accomplished by a criminal with minimal technical skill or prior experience in cybercrime.

This analysis is based on a review of one tool kit and one exploit. Although it does not necessarily represent the broader population of ransomware tool kits and actors, it supports the idea that the rapid growth of ransomware can be explained by its ability to generate value more easily, elusively, and profitably than other cybercrime-related activities. The authors also suggest several areas for further study that could alter ransomware returns on investment to the detriment of attackers. We return to this topic in the Conclusion.

4.3. Absence of Direct or Deferred Consequences

A significant financial component common in crime prevention, but absent in the fight against ransomware, is the imposition of costs after the crime. After a bank heist, for example, criminals are forced to abandon vehicles and technology. They often cannot reuse aliases that took time and money to create. They might have sunk costs in safe houses and equipment. This is often not the case in cybercrime, particularly ransomware crime. It significantly affects the cost side of the ledger when criminals know that their tools, networks, and well-being will be harmed because of their crimes [32].

In his Nobel Prize winning research into the economic framing of critical motivation, Gary Becker theorised that criminal decisions are made under a paradigm of *marginalism* which only takes into account the proximate costs and perceived benefits of the crime, with little regard given for the costs and benefits already experienced [33]. Further, Nagin and coauthors build on this premise and suggest that criminal motivations will be higher where they risks associated with the marginal decision are opaque [34]. From the criminal perspective, this likely makes ransomware especially lucrative.

We investigated this hypothesis, searching the literature on ransomware to determine if any research capable of determining the impact of *marginalism* on ransomware actors' perceptions of the value created by a cyberattack. Laszka and colleagues have developed a novel approach for pricing the optimum ransom demand to ensure profitability for the attacker. It highlights the lucrative opportunities for attacker revenue creation, given the current constraints of the system [35]. That said, revenue represents only one side of the equation. Profit requires the subtraction of expenses and other costs from gross revenues. Laszka et al. suggest a formulation for calculating the execution cost of ransomware attacks. The entire analysis merits consideration, but the core function posits a straightforward calculation of the unit cost of the attack, consisting of a valuation of the attacker's time plus the cost of developing or acquiring the attack software. The authors concede, however, that this issue is understudied, and while they do arrive at some interesting methods for estimating the value of the attacker's time, there was insufficient data to calculate the overall attack cost using this method at the time of the article's publication.

During our review of the topical literature, we did not identify any method that can suitably model costs and the breakeven point where commercially motivated ransomware attacks stop being profitable. There are, however, some interesting results from the private sector. Published in 2011, Martin's 'Cyber kill chain' whitepaper identified seven steps that cyber actors must take to complete an attack [36]. Briefly summarised, the steps are: (1) 'reconnaissance' to identify an exploitable target; (2) 'weaponisation' of a payload capable of exploiting the vulnerable system; (3) 'delivery' of the payload via some mechanism, i.e. phishing; (4) successfully bypassing installed controls, such as anti-virus, and 'exploiting' the victim system; (5) 'installation' of a second-stage malware with the ability to conduct the intended activity of those data without being blocked; (6) 'command and control' of the victim system by the attacker; and (7) 'actions on intent', such as key exchange and encryption for a ransomware attack. The article articulated a method for modelling an attack that allows defenders to target each step of the attacker's actions. Although some of the terminology may seem obscure, it allowed for much more complex attack vectors to be grouped for analysis and countermeasure. This led to an approach in cybersecurity, known as 'intelligence-driven defence', which has been used as the basis for numerous cybersecurity innovations. The result has increased not only the defence efficacy but also the cost of performing attacks significantly.

In a recent interview, Mike Poddò (one of the coauthors of the original ‘Cyber kill chain’ article) explained the results of a career spent applying intelligence-driven defence to deter attacks:

‘Even well-funded, professional cyber actors operate with limited resources, this includes financial resources, but also includes time, patience, and rare zero-day exploits’.⁴

Poddò goes on to explain that, by analysing an attack at all seven stages of the kill chain, he was able to prioritise controls focused on each stage. This was done to maximise protection, but there are further benefits. For example, an attack might successfully bypass controls at the first four stages only to be caught at the fifth stage. However, the attacker is often blind to where the failure occurred. They know only that the attack failed and that there was no response from the device they were attempting to infect. They would then often replace every element of the attack infrastructure used in the first five stages. Poddò speaks of regularly seeing attackers discard perfectly good command and control infrastructure (which was unknown to defenders and was not being blocked) out of fear that it may have been detected. There were also times when his team discovered rare zero-day exploits, not through research or complex modelling but because they detected the attack using conventional controls at a subsequent stage and then reverse-engineered the initial exploit. Over time, even the most well funded attackers would tire of burning resources. Poddò had the following to say about the impact of this method of defence on attacker morale:

It’s hard to know anyone’s precise motivations, but we have KPIs [key performance indicators] associated with our jobs. If you were a hacker and your job was to successfully target companies in the defense and security sectors, wouldn’t you get tired of showing reports that indicated you spent lots of hours, burned through lots of vulnerabilities and malware that were painstakingly developed, and had no successful compromises to show for it? [37]

The question is obviously rhetorical; we would likely answer it in the affirmative. The cyberworld includes endless potential targets. The experiences Poddò recounts indicate that cyber attackers are motivated to maximise the return on their investments of time and energy. It also suggests that the incentive to engage in the attack decreases as both actual and opportunity costs for an attack increase.

4——‘Zero-day vulnerability’ is an industry term used to describe vulnerabilities discovered by an attacker before the manufacturer of the software discovers them. There are then no developed patches or countermeasures in place. Once used, the vulnerability is traceable and the software manufacturer can develop fixes. The day the fixes are released is counted as day 1 of the vulnerability’s life.

5. How Should Society Respond: Effective Ethical, Social, and Legal Constraints on Ransomware

Perspectives on the ethical implications of preventing cybercrime vary. According to Hollis and Ohlin, ethical actions concerning cybercrime should align with ‘self-defense, economic interests in protecting intellectual property, and public health’ [38]. Much scrutiny has been applied to regulatory interventions targeted at cybercriminals but impacting citizen privacy as collateral damage. Critiques of these actions are numerous and are outside the scope of this article. More relevant to this review is the efficacy of these interventions at cybercrime deterrence. Here, the evidence suggests attempts to control cybercrime through purely punitive means have largely failed to keep up with the forces compelling its growth.

Law enforcement has mostly been slow to adapt rules of evidence and patterns of investigation to digital crimes [39]. Governments also struggle to deal with the transnational nature of most cybercrimes and the methodological process of international adjudication. Cyberspace facilitates borderless digital theft and hacktivism unmoored from standard constraints of proximity in the physical world. The crimes occur in a new domain of competition where there are no established norms for social pressures to act as restraints on bad behaviour [40]. Governance structures still observe Westphalian boundaries that do not apply to the digital contours of cyberspace [41].

In addition to the ambiguous and inadequate governance of cyberspace, the growth of ransomware also benefits disproportionately from advances in anonymous cryptocurrency payment mechanisms [41]. Paquet-Clouston and colleagues argue that the widespread popularity of cryptocurrencies, such as bitcoin, has made a once fraught exchange of value low-risk and largely seamless. This is somewhat unique in the exchange of stolen goods. Usually, stolen property – art objects, for example – trade at a significantly reduced value owing to potential forfeiture and penalties for trading in stolen goods. A conventional ransom exchange is especially fraught because the currency can be traced, and both the kidnappers and victims are physically vulnerable. Current governance structures and ethical pressures do not allow the imposition of the same constraints on cyber ransom.

Moreover, corporate shareholder interests are often misaligned with those of stakeholders. As Etzioni argues, a range of factors misalign the interests of corporations – typically the most

significant victims of ransomware – with societal aspirations (whether individual or collective) [42]. Etzioni states four reasons for this: concerns about cost, regulatory burden, consumer pressure, and efficacy. He illustrates with an analogy to historical self-regulation challenges regarding environmental pollution. Quoting a cybersecurity expert at the Security and Exchange Commission, Etzioni writes:

Cybersecurity resembles environmental law in that both fields are primarily concerned with negative externalities. Just as firms tend to underinvest in pollution controls because some of the costs of their emissions are borne by those who are downwind, they also tend to underinvest in cyber defenses because some costs of intrusions are externalised onto others. [42]

To address this imbalance, a combination of social pressure, criminal penalties, public policy, and financial disincentives is required. To be done with the highest degree of efficacy, a policy should align corporate, individual, and societal interests.

5.1. Relevant Literature in the Field of Criminology

Cybercrime occurs in a digital but not invisible marketplace. Many criminal cyber transactions market illicit goods deniably on the dark web and the exchange of value occurs online based on fictitious and deniable personas. Many crimes, ranging from illegal distribution of narcotics to wildlife trafficking, were once primarily confined to the terrestrial domain but now leverage the discretion of deniable cyberspace. This is especially well documented in literature on criminology as catalogued by Sebach in *Policing illegal drug and wildlife trades* [43]. Yet, although the research demonstrates that it is possible to observe the illicit trade on the dark web and apply specialised policing techniques, these have had limited affect owing considerably to the complexity of the jurisdictional environment and the lacking specialisation of law enforcement in digital forensics. Still, cybercrime is overwhelmingly conducted for profit, and law enforcement actions resulting in judicial penalty are only one means of affecting actor motivations. A range of law enforcement and adjacent organisations (some state-sanctioned and others not) have demonstrated their ability to affect criminal behaviour by raising real and perceived costs to the criminals. As Button demonstrated in *Private Policing*, the critical factor is for law enforcement actions to align with the public's perceived and real interests, not only to align against the interests of criminals [44].

We examine this through analysis of a conventional variant of a similar crime in the following section.

5.2. Conventional Kidnapping and Ransom Case Study

As mentioned, the ethical implications of ransomware-related crimes are understudied. However, analyses and evaluations of more conventional ransom-related crimes are quite robust. Consider the rise of kidnappings for ransom in Latin America in the late 20th century. Studies adopting the GELSI approach to conventional kidnapping could illuminate the ransomware problem. The National Defense University's Marks notes that the Revolutionary Armed Forces of Colombia's (FARC) use of kidnapping as a tool to generate ransom-related revenues in the 1980s and 1990s progressed from a source of minor revenue to the primary means of operational finance [45]. From a governmental and ethical perspective, this was considered far more benign than FARC's narcotics activity or its violent campaign against the government. It also furthered the local perception of FARC members as freedom fighters. Funds were extracted from wealthy foreign corporations, many of which were viewed by working-class locals as exploiting the country. Violence was also generally directed at foreigners, and most kidnappingees were eventually returned alive.

This coincided with the mainstreaming of Kidnap and Ransom (K&R) insurance, offered primarily to expatriate executives from the United States and Europe. Ransom payment generally resulted in favourable outcomes. Nonetheless, some evidence suggests that this also created a moral hazard. The presence of insurance contracts and the likelihood of seamless high-value payouts caused what is known in the insurance industry as *adverse selection*: being insured increases the risk of kidnapping [46]. Kidnappings in Colombia rose from 42 in 1982 to 3572 per year by the end of the century, an increase of more than 8000% [47]. By the early 1990s, Colombia had grown to lead the world in kidnappings. K&R insurers were quick to recognise this trend. They responded with a series of requirements for new insurance policies that effectively reduced adverse selection effects. Payouts to groups, such as FARC, also decreased because the US Office of Foreign Asset Control (OFAC) employed an international governance approach focused on terrorist financing. Partner nations adopted similar methods [48]. Other political and social factors likely co-contributed to reducing kidnappings in Colombia. These are addressed by Pires et al. and it is informative to read their conclusion in its entirety [47]. However, there appears to be a clear correlation between measures taken by

insurers and regulators on the one hand and decreasing kidnappings on the other. By 2010, the overall frequency of kidnapping for ransom in Colombia had dropped by 91% [47].

5.3. Adverse Selection and Moral Hazard in Responses to Ransomware

The cautionary tale of the K&R insurance market is illustrative of the present-day dynamics in cyber and ransomware insurance. We see a similar adverse selection bias in ransomware activities. The present cyber insurance market appears to be driven by the rise in ransomware targeting commercial enterprises. However, cyber insurance also contributes to the sharp ransomware growth curve. Baker and Shortland, reflecting on the previously mentioned ransomware incident at Colonial Pipeline, noted that insurance may have contributed to a double failure, first failing to incent Colonial to achieve a security posture capable of limiting the damage of the hack and then by paying a large and public ransom that likely incented other bad actors [49]. According to Manky from cybersecurity company Fortinet, ransomware attackers will search a victim network for evidence of ransomware insurance contracts [50]. The attackers often take a particular interest in the deductible and maximum payouts guaranteed by a policy. We also see a trend in pricing related to the requested ransom that closely tracks conventional kidnapping and ransom. Attackers frequently align the ransom amount with their understanding of typical ransomware coverage to maximise returns and expedite payment [41].

There are then evident similarities between the two ransom- and extortion-based insurance markets. Just as abuse of K&R insurance led to hardening of industry standards for security, Mott et al. demonstrate the sharp increase in ransomware crimes led to the insurance industry putting significant pressure on companies to improve internal security controls before they would be deemed 'insurable' [51]. There also appear to be similarities between the decision calculus of those paying the ransom. Connolly and Hervé reflect on more than 40 specific ransomware cases and document that, even when benefit of payment appears clear, the victims considered a range of views about ethics of rewarding the attacker or the degree to which they could trust their guarantees, each making their decision far more complex [52]. However, it remains unclear whether governance measures targeted at reducing incentives for payment will result in similar reductions in ransomware attack frequency. This area demands further study, as the point is less obvious than it may seem. On the one hand, it is reasonable to expect

that governance of payout mechanisms and checks on adverse selection effects will drive ransomware attack numbers down. On the other hand, there are notable dissimilarities between the two cases that may directly affect the efficacy of such controls and require careful investigation.

For example, efforts have been made to reduce payments to ransomware actors through governance actions, such as OFAC. Some ransomware cartels have been labelled terrorist organisations, and ransom payments compared to terrorist financing. This mirrors the designation given to FARC in Colombia. However, unlike the Colombian example, the actor-and-victim relationship is not geographically bound when it comes to cybercrime. In Colombia, the actors were members of a known group that physically congregated, organised in camps within FARC-controlled territory, and considered themselves members of an organisation with rank and hierarchy. Such a group can be designated an identifiable terrorist organisation and/or added to a banned list [46]. However, in the case of cybercrimes, attribution is non-geographical and often beyond the technical means of the victims. Cartel members may be distributed worldwide, and group affiliation may be discrete. Misattribution of attacks by ransomware syndicates known to be on a banned list will likely diminish the effect of these measures. A detailed investigation of the mechanisms that would disincentivise ransomware attacks is sorely needed but would be far from simple.

6. Conclusion

The frequency and extent of the damage continue to grow. The actor rationale behind this growth is straightforward: ransomware simplifies the attacker value chain. It commoditises the victim's data, selling access to such data back to the victim. It exploits vulnerabilities that are abundantly available in software and computing systems. The illicit market for ransomware tool kits and exploits operates efficiently, where the most powerful malware and prolific actors rise to the top. This market is widely accessible to parties with a range of technical skills. It offers attack building blocks and raw materials to the technologically adept; it offers 'ransomware as a service' for the technophobe. The barriers to entry are low, and the return is high (and growing higher). What further conclusions can be drawn from this realisation? Are there areas of investment or study that could alter the current incentive model, thus forcing the curve of ransomware growth downward?

It is clear that ransomware actors operate with motivations similar to those of other more conventional criminal actors. As a result, it stands to reason that policies targeted at their motivations would likely have a limiting effect. This was demonstrated by Waldrop in 2016, chronicling a series of punitive efforts directed at cybercriminals. They found that a law enforcement takedown of one group might have a ‘creative destruction’ affect similar to the failure of a business in the conventional economy, but also that punitive efforts that raised the cost of a material need by criminals did impact their behaviour, driving it away from the cost increase and towards an alternative [53]. Yet, it remains true that the overwhelming majority of the academic literature focuses on the technical nature of ransomware crimes. Our research found that the majority of hard data on attacker activity, motivations, and transactions comes from industry. Only a handful of sources addressed the multi-disciplinary *who* and *why* questions that were our scope for this review.

Still, from the literature that does exist, it seems clear that reducing the financial benefits would significantly reduce the frequency of ransomware crimes, given that ransomware actors are primarily motivated by monetary gain. Given Schneier’s observation that the majority of cybercriminals are low skill and low focus, combined with Hill’s [24] observation of the low average individual return, simply raising the real or opportunity costs of carrying out ransomware attacks could significantly reduce the frequency of ransomware crimes. Furthermore, concentrated social and legal pressure applied against the comparatively small number of criminal cartels generating disproportionate harm could have an outsized impact on the value realised by these organised ransomware actors owing to the concentration of the ransomware market (as measured by HHI). The economic impact on corporations and the life-threatening implications for individuals should motivate further innovations to reduce ransomware incentives. If properly understood, this could have the effect of leading to a greater convergence between societal norms and social values in cyberspace that might disincentivise criminal behaviour and lead to a greater degree of public diligence and corporate compliance. They could drive general acceptance of business models for technology products that impose a modest amount of friction for consumers, but with the benefit of rendering criminal technology business models obsolete.

There are many examples of similar parallels emerging in society and governance. Consumers first sought optional safety features in vehicles, many of which became standards enshrined in transportation regulations. Were the standards and governance removed,

it is unlikely that consumers would readily go back to driving vehicles without seat belts, airbags, or anti-lock brakes – suggesting that the features provide value that exceeds the mandated compliance. Individuals also readily accept a slight delay in access to their funds from the banking system to allow for transactions to clear, so as to reduce the risk of fraudulent transactions. It is certainly conceivable that similar concessions could be made in cyberspace if it was clearly demonstrated that the cost to society was well below the cost imposed on criminal actors. It stands to reason that such innovations would fundamentally reduce the real and perceived value of financially motivated cybercrime.

It is the conclusion of this review that cyber governance strategies that address the growth of cybercrime in general, and ransomware specifically, are understudied and badly needed. Further research needs to be done on how to provide potential victims and societies with significant leverage against attackers. Some limited work in this direction has begun [54], providing an excellent starting point. However, if society is to successfully combat cybercrime, effective governance must consider the social and financial costs of remedies and ensure that the costs are aligned with societal norms and values with the costs primarily allocated to the bad actors. A detailed study of these costs, both allocated to society and to cybercriminal, is necessary. It should engage the domains of economics and criminology to the same or greater degree than that of computer science, and should focus on demonstrating specifically the point at which social, legal, and financial pressures can bring the cost of conducting ransomware attacks equal to the value likely to be achieved by the cybercriminal. Such a study lies beyond the scope of this review article, but is planned as the topic of forthcoming research.

References

- [1] VirusTotal Blog. (2021). *Ransomware in a global context*. [Online]. Available: <https://blog.virustotal.com/2021/10/ransomware-in-global-context.html> [Accessed: Feb. 03, 2024].
- [2] Verizon Business. (2021). *DBIR results & analysis*. [Online]. Available: <https://www.verizon.com/business/resources/reports/dbir/2021/results-and-analysis/> [Accessed: Apr. 01, 2023].
- [3] Businesswire. (Aug. 12, 2021). *IDC survey finds more than one third of organizations worldwide have experienced a ransomware attack or breach*. [Online]. Available: <https://www.businesswire.com/news/home/20210812005739/en/IDC-Survey-Finds-More-Than-One-Third-of-Organizations-Worldwide-Have-Experienced-a-Ransomware-Attack-or-Breach> [Accessed: Apr. 01, 2023].

- [4] TechTarget. (2023). *Ransomware trends, statistics and facts in 2023*, Security. [Online]. Available: <https://www.techtarget.com/searchsecurity/feature/Ransomware-trends-statistics-and-facts> [Accessed: Apr. 01, 2023].
- [5] C. Radu, N. Smaili, "Board gender diversity and corporate response to cyber risk: Evidence from cybersecurity related disclosure," *Journal of Business Ethics*, vol. 177, no. 2, pp. 351–374, 2021, doi: [10.1007/s10551-020-04717-9](https://doi.org/10.1007/s10551-020-04717-9).
- [6] 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.
- [7] GlobeNewswire Inc. (2020). *Cybercrime to cost the world \$10.5 trillion annually by 2025*. [Online]. Available: <https://www.globenewswire.com/news-release/2020/11/18/2129432/0/en/Cybercrime-To-Cost-The-World-10-5-Trillion-Annually-By-2025.html> [Accessed: Oct. 09, 2023].
- [8] R. Anderson, C. Barton, R. Bohme, R. Clayton, C. Gañán, T. Grasso, M. Levi, T. Moore, M. Vasek, "Measuring the changing cost of cybercrime," in *Proceedings of the 18th Workshop on the Economics of Information Security (WEIS)*, Dec. 2018.
- [9] T. Seals. (2020). "FIN11 cybercrime gang shifts tactics to double-extortion ransomware," *The Cybersecurity Review*. [Online]. Available: <https://www.cybersecurity-review.com/news-october-2020/fin11-cybercrime-gang-shifts-tactics-to-double-extortion-ransomware/> [Accessed: Apr. 01, 2023].
- [10] M. Novinson. (Dec. 23, 2021). "The 10 biggest cyber and ransomware attacks of 2021," *CRN.com*. [Online]. Available: <https://www.crn.com/slide-shows/security/the-10-biggest-cyber-and-ransomware-attacks-of-2021> [Accessed: Aug. 27, 2024].
- [11] M.J. Grant, A. Booth, "A typology of reviews: an analysis of 14 review types and associated methodologies," *Health Information & Libraries Journal*, vol. 26, no. 2, pp. 91–108, 2009, doi: [10.1111/j.1471-1842.2009.00848.x](https://doi.org/10.1111/j.1471-1842.2009.00848.x).
- [12] A. Kharraz, W. Robertson, D. Balzarotti, L. Bilge, E. Kirda, "Cutting the Gordian knot: A look under the hood of ransomware attacks," in *Detection of intrusions and malware, and vulnerability assessment*, DIMVA 2015. Lecture Notes in Computer Science, vol. 9148, M. Almgren, V. Gulisano, F. Maggi, Editors. Cham: Springer, 2015, doi: [10.1007/978-3-319-20550-2_1](https://doi.org/10.1007/978-3-319-20550-2_1).
- [13] T. McIntosh, A.S.M. Kayes, Y.-P.P. Chen, A. Ng, P. Watters, "Ransomware mitigation in the modern era: A comprehensive review, research challenges, and future directions," *ACM Computing Surveys*, vol. 54, no. 9, pp. 1–36, 2022, doi: [10.1145/3479393](https://doi.org/10.1145/3479393).
- [14] R. Richardson, M. North, "Ransomware: Evolution, mitigation and prevention," *International Management Review*, vol. 13, no. 1, pp. 10–21, 2017.
- [15] J. Lee, "State of Security 2024," (2024). Splunk. [Online]. Available: https://www.splunk.com/en_us/pdfs/gated/ebooks/state-of-security-2024.pdf [Accessed: Jun. 10, 2024].
- [16] Verizon Business. (2022). *DBIR report 2022 – Summary of findings*. [Online]. Available: <https://www.verizon.com/business/resources/reports/dbir/2022/summary-of-findings/> [Accessed: Jun. 10, 2024].

- [17] I. Yaqoob, E. Ahmed, M.H. Rehman, et al. "The rise of ransomware and emerging security challenges in the Internet of things," *Computer Networks*, vol. 129, no. 2, pp. 444–458, 2017, doi: [10.1016/j.comnet.2017.09.003](https://doi.org/10.1016/j.comnet.2017.09.003).
- [18] T. Mulhall "Computer-related fraud and its evolution within telephony," *Computers & Security*, vol. 16, no. 6, pp. 521–521, 1997, doi: [10.1016/S0167-4048\(97\)84676-7](https://doi.org/10.1016/S0167-4048(97)84676-7).
- [19] A. Pawlicka, M. Choraś, Pawlicki, "The stray sheep of cyberspace a.k.a. the actors who claim they break the law for the greater good," *Personal and Ubiquitous Computing*, vol. 25, no. 5, pp. 843–852, 2021, doi: [10.1007/s00779-021-01568-7](https://doi.org/10.1007/s00779-021-01568-7).
- [20] I. Harford. (Oct. 2021). "The history and evolution of ransomware," *Tech Target*. [Online]. Available: <https://www.techtarget.com/searchsecurity/feature/The-history-and-evolution-of-ransomware> [Accessed: Sep. 26, 2022].
- [21] M. Egan. (2021). "Gasoline demand spikes in several states after pipeline hack," *CNN Business*. [Online]. Available: <https://www.cnn.com/2021/05/11/business/gas-shortage-demand-pipeline-hack/index.html> [Accessed: Jun. 12, 2024].
- [22] Homeland Security Newswire. (Apr. 21, 2010). *World's youngest known hacker caught*. [Online]. Available: <https://www.homelandsecuritynewswire.com/worlds-youngest-known-hacker-caught> [Accessed: Jun. 12, 2024].
- [23] B. Gilbert. (2019). "Lamborghinis, baby lions, and stacks of cash: The Russian hackers in charge of 'Evil Corp' are living an absurdly lavish lifestyle," *Business Insider*. [Online]. Available: <https://www.businessinsider.com/millionaire-russian-hackers-evil-corp-car-pictures-video-2019-12>. [Accessed: Jun. 12, 2024].
- [24] M. Hill. (2016). "'Flipping the economics of attacks' – A report," *Infosecurity Magazine*. [Online]. Available: <https://www.infosecurity-magazine.com/blogs/flipping-the-economics-of-attacks/> [Accessed: Sep. 26, 2022].
- [25] CSO Online. (2016). *A booming business: The rise of cybergangs* [Online]. Available: <https://www.csoonline.com/article/559369/a-booming-business-the-rise-of-cybergangs.html> [Accessed: Feb. 01, 2024].
- [26] I. Thornton-Trump, "Malicious Attacks and Actors: An Examination of the Modern Cyber Criminal," *EDPACS*, vol. 57, no. 1, pp. 17–23, 2018, doi: [10.1080/07366981.2018.1432180](https://doi.org/10.1080/07366981.2018.1432180).
- [27] W. Gragido, *Blackhatonomics an inside look at the economics of cybercrime*, 1st ed. Amsterdam: Syngress, 2013.
- [28] D. Freeze. (2021). "Cybercrime to cost the world \$10.5 trillion annually by 2025," *Cybercrime Magazine* [Online]. Available: <https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/> [Accessed: Jun. 12, 2024].
- [29] Verizon Business. (2024). *DBIR report 2024 - Summary of findings*. [Online]. Available: <https://www.verizon.com/business/resources/reports/dbir/2024/summary-of-findings/> [Accessed: Jun. 12, 2024].
- [30] D.B. Cornish, R.V. Clarke, "Understanding crime displacement: An application of rational choice theory," *Criminology*, vol. 25, no. 4, pp. 933–948, 1987, doi: [10.1111/j.1745-9125.1987.tb00826.x](https://doi.org/10.1111/j.1745-9125.1987.tb00826.x).
- [31] K. Huang, M. Siegel, K. Pearson, S. Madnick, "Casting the dark web in a new light," *MIT Sloan Management Review*, vol. 61, no. 1, pp. 84–85, 2019.

- [32] The Azure Forum. (2022). *Deterring ransomware attacks: Treat ransomware as criminality* [Online]. Available: <https://www.azureforum.org/deterring-ransomware-attacks-as-an-international-security-priority-treat-ransomware-as-criminality/> [Accessed: Jun. 12, 2024].
- [33] G.S. Becker, "Crime and punishment: An economic approach," *Journal of Political Economy*, vol. 76, no. 2, p. 169, 1968, doi: [10.1086/259394](https://doi.org/10.1086/259394).
- [34] D.S. Nagin, F.T. Cullen, C.L. Jonson, Eds., *Deterrence, choice, and crime, vol. 23: Contemporary perspectives*. New York: Routledge, 2018. doi: [10.4324/9781351112710](https://doi.org/10.4324/9781351112710).
- [35] A. Laszka, S. Farhang, J. Grossklags, "On the economics of ransomware," *arXiv.org*, 2017, doi: [10.48550/arxiv.1707.06247](https://doi.org/10.48550/arxiv.1707.06247).
- [36] Martin L. (2015). *Cyber kill chain*® [Online]. Available: <https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html> [Accessed: Jun. 12, 2024].
- [37] M. Poddo, "Cyber killchain interview," Jun. 2021. Interview by J. Jaffe.
- [38] D.B. Hollis, J.D. Ohlin. (2018). *What if cyberspace were for fighting?* [Online]. Available: https://solo.bodleian.ox.ac.uk/discovery/fulldisplay?docid=cdi_gale_infotracmisc_A570532983&context=PC&vid=44OXF_INST:SOLO&lang=en&search_scope=MyInst_and_CI&adaptor=Primo%20Central&tab=Everything&query=any,contains,what%20if%20cyberspace%20was%20for%20fighting&offset=0 [Accessed: Jan. 31, 2024].
- [39] G. Gogolin, J. Jones, "Law enforcement's ability to deal with digital crime and the implications for business," *Information Security Journal*, vol. 19, no. 3, pp. 109–117, 2010, doi: [10.1080/19393555.2010.483931](https://doi.org/10.1080/19393555.2010.483931).
- [40] P. McGuinness, "Interview with Paddy McGuinness," Jan. 2022. Interview by J. Jaffe.
- [41] M. Paquet-Clouston, B. Haslhofer, B. Dupont, "Ransomware payments in the Bitcoin ecosystem," *Journal of Cybersecurity*, vol. 5, no. 1, pp. 1–11, 2019, doi: [10.1093/cybsec/tyz003](https://doi.org/10.1093/cybsec/tyz003).
- [42] A. Etzioni. "Cybersecurity in the private sector," *Issues in Science and Technology*, vol. 28, no. 1, pp. 58–62, 2011.
- [43] L. Sebagh, *Policing illegal drug and wildlife trades – The role of the police, legal online platforms, private organisations and individuals, and cybercriminal traders*. PhD thesis, Oxford: University of Oxford, 2021.
- [44] M. Button, "Voluntary policing," in *Private policing*, 2nd ed. London: Routledge, 2019, pp. 128–151. doi: [10.4324/9781351240772-8](https://doi.org/10.4324/9781351240772-8).
- [45] T.A. Marks, "FARC, 1982–2002: Criminal foundation for insurgent defeat," *Small Wars & Insurgencies*, vol. 28, no. 3, pp. 488–523, 2017, doi: [10.1080/09592318.2017.1307612](https://doi.org/10.1080/09592318.2017.1307612).
- [46] P.L. Brockett, L.L. Golden, S. Zapparoli, J.M. Lum, "Kidnap and ransom insurance: A strategically useful, often undiscussed, marketplace tool for international operations," *Risk Management and Insurance Review*, vol. 22, no. 4, pp. 421–440, 2019, doi: [10.1111/rmir.12134](https://doi.org/10.1111/rmir.12134).

- [47] S.F. Pires, R.T. Guerette, C.H. Stubbart, "The crime triangle of kidnapping for ransom incidents in Colombia, South America: A 'Litmus' test for situational crime prevention," *British Journal of Criminology*, vol. 54, no. 5, pp. 784–808, 2014, doi: [10.1093/bjc/azu044](https://doi.org/10.1093/bjc/azu044).
- [48] US Department of the Treasury. (Apr. 22, 2008). *Treasury targets FARC financial network in Colombia*. [Online]. Available: <https://home.treasury.gov/news/press-releases/hp938> [Accessed: Jun. 12, 2024].
- [49] T. Baker, A. Shortland, "Insurance and enterprise: Cyber insurance for ransomware," *Geneva Papers on Risk and Insurance – Issues and Practice*, vol. 48, no. 2, pp. 275–299, 2023, doi: [10.1057/s41288-022-00281-7](https://doi.org/10.1057/s41288-022-00281-7).
- [50] D. Manky, J. Richberg (Feb. 17, 2022). *Ransomware cyber insurance & settlements Q&A*, Fortinet Blog. [Online]. Available: <https://www.fortinet.com/blog/industry-trends/qa-ransomware-settlements-and-cyber-insurance> [Accessed: Jun. 12, 2024].
- [51] G. Mott, S. Turner, J.R.C. Nurse, J. MacColl, J. Sullivan, A. Cartwright, E. Cartwright, "Between a rock and a hard(ening) place: Cyber insurance in the ransomware era," *Computers & Security*, vol. 128, 2023, doi: [10.1016/j.cose.2023.103162](https://doi.org/10.1016/j.cose.2023.103162).
- [52] A. Yuryna Connolly, H. Borrión, "Reducing ransomware crime: Analysis of victims' payment decisions," *Computers and Security*, vol. 119, no. C, 2022. doi: [10.1016/j.cose.2022.102760](https://doi.org/10.1016/j.cose.2022.102760).
- [53] M.M. Waldrop, "How to hack the hackers: The human side of cybercrime," *Nature*, vol. 533, no. 7602, pp. 164–167, 2016, doi: [10.1038/533164a](https://doi.org/10.1038/533164a).
- [54] Verizon Business. (2023). *DBIR report 2023 – Summary of findings*. [Online]. Available: <https://www.verizon.com/business/resources/reports/dbir/2023/summary-of-findings/> [Accessed: Jun. 12, 2024].

Disjointed Cyber Warfare: Internal Conflicts among Russian Intelligence Agencies

Cosimo Melella | CCDCOE, Tallinn, Republic of Estonia, University of Genoa, Italy | ORCID: 0009-0009-6970-9396

Francesco Ferazza | Royal Holloway, University of London, UK |
ORCID 0009-0005-3280-2678

Konstantinos Mersinas | Royal Holloway, University of London, UK |
ORCID 0000-0002-4402-2987

Abstract

Our ongoing, descriptive study explores the intricacies of Offensive Cyber Operations (OCOs), particularly in the context of the Russian-Ukrainian conflict that began in 2022. This conflict has underscored an escalation in Russian cyber capabilities. Despite OCOs playing a role, academic research indicates a relatively limited 'spillover effect'. Our study aims to investigate this limited spillover, focusing on the lack of collaboration among Advanced Persistent Threat (APT) groups associated with Russian intelligence agencies: GRU, SVR, and FSB. By analysing the operational and technical integration among these agencies, we seek to identify factors influencing cooperation. Preliminary findings suggest that internal competition and historical disparities may have hindered effective coordination in cyber operations. We posit that this lack of coordination could potentially reduce cyberattack effectiveness and increase detection likelihood. Importantly, we recognise that behavioural aspects, such as the principal-agent problem, may contribute to the barriers preventing collaboration and coordination. These behavioural factors, alongside institutional rivalries, likely play a significant role in shaping the competitive dynamics among

Received: 11.06.2024

Accepted: 07.08.2024

Published: 06.09.2024

Cite this article as:

C. Melella, F. Ferazza, K. Mersinas "Disjointed Cyber Warfare: Internal Conflicts among Russian Intelligence Agencies," ACIG, vol. 3, no. 2, 2024, pp. 38-71. DOI: 10.60097/ACIG/192120

Corresponding author:

Cosimo Melella, CCDCOE, Tallin, Republic of Estonia and University of Genoa, Italy; E-mail: cosimo.melella@ccdcoe.org

 0009-0009-6970-9396

Copyright:

Some rights reserved:
Publisher NASK



Russian intelligence agencies. As our research progresses, we aim to explore the implications of this internal rivalry on the development of technical infrastructure for Russia-affiliated APT groups. We anticipate that our findings illuminate the reasons behind the apparent reduced effectiveness of cyberattacks in this scenario. This exploration of competitive dynamics, historical nuances, and behavioural factors within Russian intelligence agencies is crucial for a comprehensive understanding of the broader cyber operations landscape. We present this paper as a work in progress, aiming to contribute to the ongoing discourse in this field.

Keywords

cyber threat, intelligence, APTs, coordination, cooperation

1. Introduction

According to Damjan Štrucl, the role of Offensive Cyber Operations (OCO) in modern conflicts has been notably heightened by the Russian invasion of Ukraine on February 24, 2022. Prior analyses, drawing from precedents like Stuxnet and NotPetya, had projected a significant impact of cyber warfare, particularly through malware distribution with potential repercussions extending beyond the immediate conflict zone to affect other nations and organisations. This expectation was underpinned by the recognition of Russia's formidable cyber capabilities. Yet, the unfolding of events presented a striking puzzle: contrary to widespread predictions, the Russian OCOs manifested limited effects on the war's outcome. This discrepancy was highlighted in several assessments that questioned the anticipated dominant role of cyber operations in the conflict. On the one hand, forecasts had envisioned a scenario where cyber operations would play a pivotal role in the warfare strategy; on the other, post-event analyses and reports underscored the surprisingly marginal impact of these operations. This apparent paradox suggests a lack of coordination among Russian intelligence agencies as a plausible explanation [1]. These empirical observations introduce a theoretical quandary: How can coordination be managed or integrated within OCOs? This is a work in progress and presents an exploratory study into a complex theoretical challenge: understanding the dynamics of coordination within OCOs, particularly in the context of Russian intelligence agencies. The study identifies a crucial observation that GRU, SVR, and FSB [2] are indeed distinct entities, each operating with unique strategies, technologies, and protocols. This differentiation is not merely organisational but extends to their approach to

cyber operations. The real puzzle, as underscored by our research, lies in the evident challenge these organisations face in coordinating their activities effectively, despite their established distinctiveness. This lack of coordination presents a significant inquiry into why these entities, known for their respective capabilities, do not achieve a unified and cohesive operational front in cyber warfare. A key observation driving this inquiry is the apparent limited effectiveness of Russian OCOs, attributed primarily to a shortfall in operational and technical integration among these agencies. This lack of coordination, especially among various advanced persistent threats (APTs), forms the central theme of our investigation. Our approach to exploring this issue is two-fold. Initially, we delve into the notion of integration at both technical and operational levels within intelligence agencies active in cyber defence. Subsequently, we empirically analyse this concept within the framework of Russia's intelligence system. This analysis aims to illuminate the roles of internal competition and political rivalry among these agencies and how these factors might influence state-sponsored cyber threats [3]. This paper aims to contribute to the broader debate on state-sponsored cyber operations. By focusing on the possible reasons for the observed lack of coordination among different hacking groups purportedly connected to Russia, the study offers insights into the impact of internal dynamics – such as competition and rivalry within the Russian government and intelligence sectors – on the nature and structure of state-affiliated cyber threats. This perspective is novel and adds a valuable dimension to our understanding of state-sponsored cyber activities. In some cases, political rivalry can lead to a politicisation of these agencies, where officers or civil servants are chosen based on their political affiliation, rather than their qualifications or experience [4]. Such a situation can lead to deterioration in the quality of the agency's services and less trust in government institutions by the public. Collectively, political rivalry can create significant externalities [5] in the competition between public agencies, creating challenges for leaders and executive officials as they seek to deal with changing priorities while maintaining the integrity and effectiveness of their operations. In recent years, acknowledging the historical backdrop of inter-agency rivalry in Russia, particularly between the FSB and the GRU, sheds light on the complexities of coordination within its intelligence framework. Incidents such as GRU's involvement in the 2014 Crimea annexation and the handling of Sergei Skripal's poisoning in 2018 have highlighted this friction, with the FSB expressing dissatisfaction over perceived oversteps by GRU. This longstanding political rivalry among Russia's intelligence entities, including the SVR, prior to the 2022 Ukraine conflict, suggests that the observed

lack of coordination and integration during the war was, in retrospect, an anticipated outcome. Consequently, the initial expectations of a significant cyber offensive impact, akin to spillover effects seen in previous global cyber incidents, may have overlooked the practical implications of these internal dynamics, thereby contributing to the re-evaluation of the puzzle surrounding Russia's cyber operations effectiveness [6]. The landscape of inter-agency competition, compounded by political rivalries, is full of challenges. This unstable dynamic environment can induce uncertainty and instability, hampering operational and strategic coordination. To illustrate this environment, tension has been observed within the Russian intelligence community, particularly between the FSB and the GRU, due to alleged excesses of jurisdiction and operational abuse. This study adds to the talk of state-sponsored cyber operations by providing an explanatory lens for coordination deficiencies observed among hacking groups allegedly linked to Russia [7]. Furthermore, we seek to answer two central research questions (RQs) regarding the degree of integration between cyber defence agencies' operational and technical/tactical levels and the factors contributing to any observed lack of integration:

- RQ1: To what extent does integration occur between the technical and operational divisions within intelligence agencies when executing government-offensive policies in cyberspace?
- RQ2: What factors impede the integration between technical and operational divisions within intelligence agencies in the implementation of government-offense strategies in cyberspace?

In doing so, we emphasise the critical role of the technical and operational levels within intelligence agencies. While the technical level focuses on the skilful use of information management technologies, the operational level primarily addresses the strategic use of information for immediate decision-making. These two layers, while distinct, often need to be closely integrated for an effective response to threat or opportunity. Lack of coordination can lead to a significant disconnect between strategic objectives and their operational execution. This disjunction often stems from the divergence between technical capabilities and operational planning – wherein the technological approaches do not align with operational plans. Such misalignment threatens to widen the gap between what is strategically decided and what is practically implemented, resulting in technical inefficiencies, leading to operational inefficacies [8]. Our research aims to illuminate these coordination challenges and propose mechanisms for greater integration within state-sponsored cyber operations. Indeed, moving forward, let's examine the

potential implications of a fragmented intelligence community. It erodes the quality of services rendered by agencies. Furthermore, the well-known political competition within public agencies can produce significant externalities. Navigating the shifting currents of rivalries and evolving strategic priorities pose significant challenges for agency leaders and officers, potentially disrupting the effectiveness and integrity of their operations. Historical tensions within the Russian intelligence community have often led to strategic misalignments. For example, the FSB has reportedly expressed dissatisfaction with GRU's role in the 2014 annexation of Crimea, considering it a violation of its jurisdiction. Similarly, the handling of Sergei Skripal's poisoning in 2018 is said to have intensified friction between the agencies [9]. The misalignment between strategic objectives and their execution due to internal fragmentation can lead to operational inefficiencies and potential vulnerabilities, highlighting the need for better integration at technical and operational levels. We hope to contribute to the broader discourse on offensive state-sponsored cyber operations through this lens. The methodology used to answer RQs and better understand such operations is multifaceted, in the following order:

- We conduct a literature review on cyber operations, intelligence agency structures, and inter-agency dynamics.
- We analyse open-source intelligence (OSINT) data related to Russian cyber activities during the Ukraine conflict.
- We employ a case study approach, examining the activities of three main Russian intelligence agencies: GRU, SVR, and FSB, along with their associated APTs.
- We analyse the tactics, techniques, and procedures (TTPs) of specific APTs linked to these agencies, such as Sandworm, Fancy Bear, Cozy Bear, Turla, Callisto, and Gamaredon.

2. The Challenges of Coordination

In the complex landscape of OCO, the effective management of challenges heavily relies on the robust establishment of cooperation and coordination principles. Cooperation refers to sharing resources, information, or skills to achieve common goals or tackle shared challenges. Coordination refers to the organisation of the efforts of the various actors, aimed at ensuring the efficient and effective achievement of the shared objectives. At the strategic level, which involves long-term planning and decision-making aimed at achieving overarching goals, cooperation is the key. It involves a concerted effort among various organisations and entities, bridging their resources and capabilities. This level of

operation is crucial in conflict situations, requiring not just strong political determination but also a unified strategic vision to address broad, often long-term objectives. In contrast, coordination is critical at both operational and technical levels. The operational level refers to the execution of strategies, focusing on how different components of an organisation or entities work together to implement the strategic plan. This might involve day-to-day management of resources, decision-making regarding specific cyber operations, and real-time responses to evolving situations. The technical level, on the other hand, delves into the specificities of cyber warfare, dealing with the actual tools, tactics, and procedures used in cyber operations. It includes hands-on tasks, such as software development, system penetration, data analysis, and other technical aspects of cyber warfare. Coordination at this level ensures that the technical actions align with the strategic objectives and operational plans. It involves synchronising cyber operations, sharing crucial intelligence, and modifying tactics and techniques as needed to effectively counteract adversaries' defensive measures or react to their coordinated activities on the battlefield. Understanding and integrating these levels of operation is essential in managing the dynamic and intricate nature of cyber conflicts and the activities of APTs. Such an integrated approach ensures that strategic decisions are effectively translated into operational success and technical precision, a critical factor in the domain of OCOs. Referring to what has been written about the importance of coordination in OCO, the academic studies of McNeil [10], Hernandez-Ardieta, Tapiador, Suarez-Tangil [11], Heuvel, Baltink [12], and Liebetrau [13] provide further insights into this essentiality of coordination in cyberspace. These academic works reinforce the idea that to successfully face the challenges of cyberspace and effectively manage cyber operations; it is fundamental to establish solid principles of cooperation at the strategic level and coordination at all levels: strategic, operational, and technical. McNeil highlights the need for strategic international cooperation, emphasising how its absence can limit offensive and defensive capabilities in cyberspace. It reflects the importance of lower-level coordination among nations to achieve long-term objectives. The article by Hernandez-Ardieta, Tapiador, and Suarez-Tangil sheds light on the importance of information-sharing models for coordinated cyber defence, recognising the essentiality of coordination at the operational and technical levels to ensure alignment between technical actions and strategic objectives. Finally, Liebetrau, in his article, examines how different countries organise their cyber capabilities, identifying various organisational models and emphasising the importance of coordination between military and intelligence

entities, which is essential for addressing cyber conflicts. These studies emphasise that information sharing and coordination are crucial for improving operational capabilities and security in cyberspace. They highlight the importance of continuous efforts to develop effective frameworks, agreements, and protocols, ensuring that strategic decisions are translated into operational success and technical precision in OCOs. Coordination between different APTs in achieving similar or different goals depends on the goals set by their coordinating intelligence agencies. If the intent is to maximise the impact of an operation, it may be appropriate to aim simultaneously at the same goal [14]. Conversely, if the operation is aimed at stealth, cyber-espionage, or evasion of detection, it is more appropriate to target different targets simultaneously [15]. Mandiant, which has been monitoring cyber threat intelligence activities in various Ukrainian organisations since the beginning of the conflict, has reported incidents where the detection of one APT's operation led to the discovery of another APT's activities. It occurs due to data collected by Security Information and Event Management (SIEM) systems that identify specific TTPs linked to one or more threat actors. Additionally, coordination between APTs can be challenging, as it requires high trust and synergy between sponsoring organisations. This increased interaction can increase the risk of exposure and compromise, negatively affecting the operation's success. The coordination between APTs and the achievement of similar or different objectives will depend on several factors, including the operation's objectives, the resources available to the sponsoring organisations, and the target infrastructure's security posture [16]. A case in point of this scenario is the Democratic National Committee (DNC) hack in 2016, which involved two separate Russian hacker groups: APT28, affiliated with the GRU, and APT29, linked to the SVR. This cyber breach was notable for its sophistication and volume of sensitive data stolen, including emails and other DNC documents [17]. While APT28 and APT29 are commonly believed to have coordinated the hack, evidence suggests they still needed to synchronise their efforts. For example, APT28 used a spear phishing campaign to access the DNC's email system, while APT29 used a different method involving a compromised VPN. Furthermore, the tools and TTPs used by the two groups varied, indicating a target-based fit. For example, APT28 reportedly used X-Agent for data exfiltration, while APT29 used a different tool, SeaDaddy. Despite the lack of coordination, APT28 and APT29 successfully executed a cyberattack on the DNC. However, this lack of coordination may have led to overlooked opportunities or inefficiencies [18]. In recent decades, and before the invasion of Ukraine, Russia has leveraged sophisticated cyber capabilities to conduct

global disinformation campaigns, propaganda, espionage, and destructive cyberattacks. Russia oversees numerous units that carry out these operations under various security and intelligence agencies. These Russian security agencies often compete and conduct parallel operations on the same targets, complicating specific attribution assessments. Over the past two decades, Russia has expanded the staffing of its security agencies, thereby developing extensive capabilities to undertake a wide range of cyber operations. No single Russian security or intelligence agency holds sole responsibility for cyber operations. Instead, three agencies share this role: GRU, SVR, and FSB [19]. The distribution of responsibilities between GRU, SVR, and FSB can sometimes lead to overlapping or conflicting operations. Each of these agencies maintains its information units and strategic goals, which reflect the broader goals of their parent organisations. The GRU is traditionally associated with military intelligence and has been implicated in numerous cyber operations to disrupt or destabilise foreign infrastructure. It includes the DNC hack attributed to APT28, which was aligned with the GRU's more aggressive operational stance. Meanwhile, the SVR focuses on traditional espionage and foreign intelligence gathering. SVR-related cyber operations, such as those attributed to the APT29, usually reflect this goal, targeting foreign governments, organisations, and individuals for intelligence gathering, rather than disruption. Finally, the FSB, primarily an internal security agency, is also involved in cyber operations. These operations often have a more defensive slant, focusing on internal security, counter-intelligence, and maintaining control over Russia's information space. However, the FSB has also been associated with OCOs, particularly those targeting dissidents, activists, and other alleged threats to Putin's government. The division of cyber responsibilities among these agencies reflects Russia's cyber strategy's complex and multifaceted nature. However, as has been noted, this division can lead to inefficiencies and missed opportunities due to a lack of coordination. For example, the different methods and tools used by APT28 and APT29 in the DNC hack could have allowed for a more thorough or effective operation if there had been more collaboration between the two groups. While there is no indication that the GRU, SVR, or FSB will have sole responsibility for these operations, there may be increased efforts to coordinate and streamline activities between these agencies. It could lead to a more unified and powerful Russian cyber threat. However, the inherent challenges of coordinating between large and complex organisations with differing goals and operating cultures should not be underestimated [20]. A brief graphical representation of this section is shown in Figure 1.

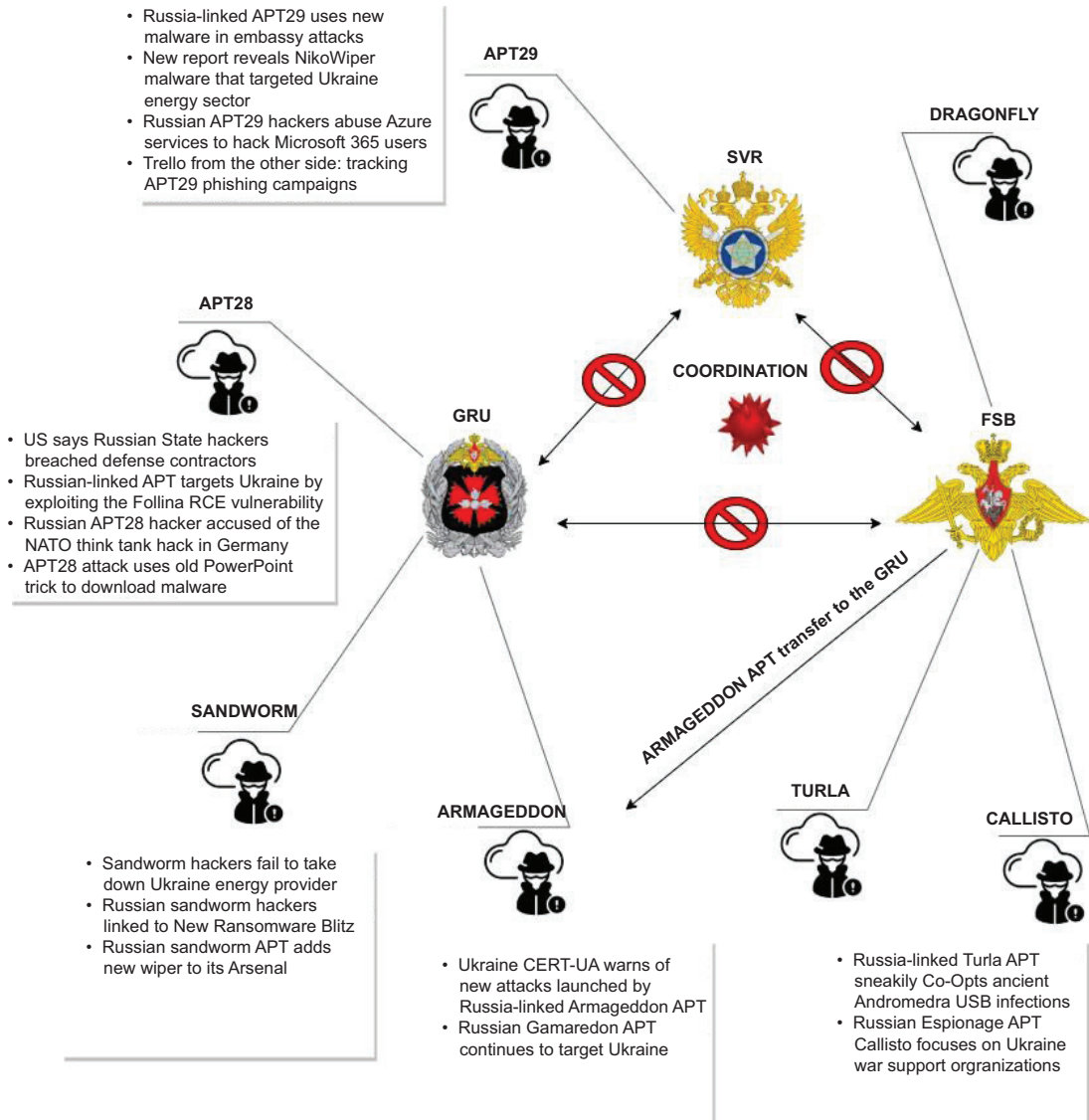


Figure 1. Coordination amongst Russian intelligence agencies and related APTs.

3. Factors Impacting Coordination

Coordination between technical and operational layers in cyberspace faces several challenges affecting the efficiency, security, and reliability of communication and collaboration. Firstly, different systems, platforms, and protocols can make seamless communication and coordination difficult. Ensuring interoperability between various devices, applications, and networks so that they work together requires standardisation, implementing standard

protocols and constant updating. Communication delays can hinder real-time coordination, especially in cases where an immediate response is needed. Latency, an additional factor, can be caused by network congestion, physical distance, or routing inefficiencies. Finally, scalability also has a direct effect. As the number of devices, users, and systems involved in cyberspace increases, ensuring that the infrastructure of one or more agencies can handle this growth becomes a challenge. Scalability issues can lead to degraded performance or even system failure [21]. Furthermore, for the above reasons, coordination fails between intelligence agencies in cyberspace (for offensive or defensive purposes [22]). The lack of coordination between the operational and technical/tactical layers of these organisations can make it more challenging to carry out attacks with a destructive effect. The lack of coordination between operational and strategic levels among cyber threat groups can lead to counterproductive outcomes, significantly hampering their collective effectiveness. When these layers fail to share information and align their efforts, they risk not only diminishing the impact of their operations but also increasing the likelihood of detection by cyber security defences. This misalignment can result in operational redundancies, conflicting actions, undermining the overarching objectives of the cyber campaign. To enhance operational security and effectiveness, establishing robust communication channels and coordination mechanisms is essential, ensuring that all actions are synergistic and strategically aligned. Cultural and historical differences between these agencies hinder effective communication and coordination in cyberspace. Added to this are confidentiality issues: the need to balance security and privacy with the ability to coordinate and share information creates technical limitations. This competition creates disjointed efforts, undermining the efficiency of cyberattacks. Intelligence agencies, rather than pursuing large-scale destructive attacks through their units, have preferred to use their APTs mainly for cyber-espionage purposes, sometimes trying to integrate the cybernetic plan with the kinetic one to achieve their operational goals [23]. Cyber operations conducted by different intelligence agencies involve a complex set of technical and operational layers working together. The technical level typically involves using advanced technologies, such as malware, remote access tools, and other sophisticated hacking techniques, to gain unauthorised access to targeted computer systems and networks [24]. Especially, cyber espionage operations conducted by different intelligence agencies involve a complex set of technical and operational layers working together. The technical level typically involves using advanced technologies, such as malware, remote access tools, and other sophisticated hacking techniques,

to gain unauthorised access to targeted computer systems and networks. The operational level, on the other hand, encompasses the execution of the operations themselves. This level involves identifying and prioritising targets, choosing appropriate methods of attack, and coordinating the actions of operators engaged in the operation. To effectively integrate the technical and operational levels, an intelligence agency typically employs highly trained agents trained to understand cyber espionage's technical and operational aspects. These operators work together in a coordinated way to develop and execute complex attacks on targeted systems and networks [25]. On a technical level, the operators use various tools and techniques to gain unauthorised access to the target's computer systems and networks. It can involve exploiting vulnerabilities in software, using phishing attacks to trick users into giving up their login credentials or using social engineering techniques to gain access to sensitive information. Once access is gained, agents can use various information-gathering tools, such as key logging software, to capture passwords and other sensitive information or malware to monitor the target's activities and communications [26]. At the operational level, operators use their understanding of target motivations and behaviour to leverage the information gathered to deploy attack tactics. For example, they can use the information to influence the target's decisions or to gather more information about other targets. Successful cyber espionage operations require high technical and tactical sophistication and a deep understanding of the target's motivations, behaviours, and vulnerabilities. The integration of technical and operational levels is essential for the success of these operations and requires a high degree of skill and coordination between the operators involved.

4. Objectives, Skills, and Culture as Coordination Challenges

While intra-agency coordination remains achievable despite challenges in melding technical and operational levels, inter-agency collaboration presents a more complex scenario due to divergent organisational cultures, conflicting priorities, infrastructural disjunctions, and varying degrees of technical and operational expertise. These dynamics underscore the need for refined RQs that capture both internal and external integration facets within intelligence agencies' cyber operations. Thus, we propose an updated framing of RQs:

- RQ1: To what extent does integration occur between the technical and operational divisions within intelligence agencies, specifically

when executing government-offensive policies in cyberspace? This question aims to explore the depth and effectiveness of internal coordination efforts, shedding light on the synergy between technological innovations and operational strategies.

- RQ2: What factors impede the integration between technical and operational divisions within intelligence agencies, particularly in the context of implementing government-defensive strategies in cyberspace? This inquiry seeks to identify and analyse the barriers to effective collaboration, focusing on the internal dynamics that challenge the alignment of cyber defence objectives with operational execution. These updated questions aim to provide a comprehensive understanding of both internal integration within agencies and complexities of inter-agency cooperation, reflecting the multifaceted nature of cyber operations in the intelligence community [27]. A key challenge is that different intelligence agencies may have different goals and priorities. For example, one agency might focus on gathering information about a particular target, while another might be more interested in disrupting the target's activities or using intelligence to influence decisions [28]. These differing priorities can make it difficult to coordinate operations effectively, as each agency may have a different approach to intelligence collection and use. In some cases, agencies may even have conflicting goals, such as when two agencies are interested in a particular target audience but have different goals and *modus operandi* on how to approach the task [29]. Another challenge is that different agencies may have different technical and operational expertise levels. For example, one agency may be more proficient at developing and executing complex cyberattacks. At the same time, another may have skill sets for gathering information from various sources and deploying psychological operations [30].

5. The Principal-Agent Dynamic

Furthermore, there may be a disruption in the principal-agent dynamic between the technical/tactical and operational levels between APTs working for different intelligence agencies and the decision-makers who deal with high-level coordination activities. The 'principal-agent problem' in economics models the situation where one or more 'agents' operate on behalf of the 'principal' who has hierarchical dominance over the agents. This relationship involves information asymmetries, since the agents usually have access to more information than the principal, and conflicts of interest, since agents might not operate in accordance with the principal's benefit. Principals cannot monitor closely the actions

of the agents, and agents have motivations which might not serve the principal's goals. In our case, conflicts can arise by a need for more understanding: actors with technical expertise working within groups may need to understand decision-makers' broader goals and strategies clearly. On the other hand, decision-makers may need help for understanding the technicalities. Furthermore, this is why decision-makers (at the strategic level) and those who execute these decisions (at the operational level), both essential elements of tactical planning, need to spend more time identifying and prioritising their goals. The problem of information sharing in this context is aggravating: intelligence agencies (acting as 'agents') have access to more information and are often reluctant to share this information with those working at the coordination level (the 'principals') or with other engineers from different entities, resulting in a lack of coordination and collaboration. Intelligence agencies may be reluctant to share information for various reasons, such as protecting sources. Disclosure of this information could put these sources or specific operations at risk. Similarly, agencies may want to protect the specific methods by which they conduct operations and collect information. If these techniques become public knowledge, they may become less effective. These bodies may want to maintain control over the information they collect to ensure it is used appropriately and to have a bargaining edge when influencing political decisions. Additionally, there may be some resistance to information sharing if agencies feel they need more recognition for their work or are concerned that other agencies may use the information to advance their interests at their own expense. These problems can lead to hampering the overall effectiveness of the intelligence system. Moreover, the principals, that is, the agency-coordinating entities at the higher level, do not necessarily share their broader strategy with the agents, that is, the agencies. Thus, in lack of a 'broader picture' (another information asymmetry), the aforementioned factors and coordinating challenges can be maintained and perpetuated. Even in the case of minimisation of information asymmetries, the historical analysis of the agencies under examination reveals an often competitive stance amongst the agencies. Whether this is a deliberately cultivated environment from senior leadership or a phenomenon that has evolved organically amongst the agencies can be debatable. But, in either way, such an environment maintains the aforementioned challenges. These differences in expertise and access to information can make it difficult to coordinate operations effectively, as agencies may need to fully understand each other's capabilities, limitations, and motivations. This setting can lead to misunderstandings or communication problems, compromising operational success. In summary,

the principal-agent dynamic highlights significant coordination and information-sharing challenges within and between intelligence agencies operating APTs. These challenges stem from information asymmetries and conflicting interests, where technical teams may lack insight into broader strategic goals and decision-makers may not grasp operational technicalities. Such disparities hinder effective cooperation and can compromise operational success. Overcoming these obstacles requires improved communication, mutual understanding of goals and methodologies, and a commitment to aligning actions with overarching strategic objectives.

6. Cultural Differences

Different organisational cultures exhibit varying behaviours and approaches; these differences might make it difficult for different intelligence agencies to work together effectively. There are several studies on the effects of cultural characteristics. Empirical research identifies a number of cultural dimensions to describe a national or regional culture. Such dimensions can be equally applied to organisations, and, for our purposes, can indicate how differences in these dimensions can impair coordination between them. While there are many of these dimensions, proposed by different researchers [31, 32], we focus on a selected subset, that is, the ones that are likely to have the highest impact on the coordination between the examined agencies. For our purposes, we consider intelligence agencies as entities which have their own characteristics, that is, they have measurable ‘scores’ across the following dimensions. One of the most relevant dimensions, in this sense, is that which describes how trust is gained, for trust is a pivotal aspect of highly confidential environments. Different organisational cultures might have different ways to attribute trust, and coordinating groups where trust is gained in different ways can be tricky. For example, one group might find higher trust value in personal relations, such as simply having attended the same military academy (relationship-based trust), while the other group might find higher trust in performance, or a long successful career with achievements (task-based trust). Another important cultural aspect is that of leadership; some organisations might be more hierarchically structured, with strict and well-defined vertically ordered ranks, while others might have more loose, egalitarian structures which reach decisions via consensus. The degree of uncertainty avoidance that an organisation can tolerate is also a very important dimension. Some organisations require everything to be normed, and deviation from these norms is often a cause of ‘neuroticism’, conflict, and confusion. Other organisations might be more flexible, being less focused on inflexible principles,

and more open to opportunity and change. Last, but not least, another relevant cultural aspect is that of decision-making; some organisations might favour a top-down approach, where leading individuals make decisions and impose these to subordinates, while others take a consensus-based approach. In the light of the above, it appears that motivations and access to information of agency entities in the form of principal-agent dynamics, or cultural differences between agencies, can amplify or diminish coordination challenges between agencies. In the next section, we present the case studies of GRU, SVR, and FSB, along with their indicative corresponding APTs. The choice to focus on GRU, SVR, and FSB agencies for the case study portion of our OCO study was driven by several significant factors. Firstly, the context of the recent Russian-Ukrainian conflict at the centre of this paper, which has seen a marked increase in Russian cyber capabilities, makes these agencies particularly relevant. The GRU, SVR, and FSB have been protagonists in various cyber operations in this context. These agencies have distinct but complementary roles in intelligence and cyber operations. The GRU deals primarily with military intelligence, the SVR with foreign intelligence, and the FSB with internal security and counter-intelligence. By analysing the interactions between these agencies, we can gain greater insight into Russia's internal dynamics in cyber operations. Another critical aspect is the historic competition and disparities between these agencies. These internal differences offer a rich context for exploring how they influence coordination and effectiveness in cyber operations. Understanding the causes of their lack of coordination can reveal key factors that hinder or facilitate greater cooperation. Furthermore, our analysis focuses on the impact of this lack of coordination on the effectiveness of cyber operations. If these agencies fail to coordinate effectively, this could reduce the impact of their cyberattacks and increase the likelihood of detection. By examining interactions at operational and technical levels, our study seeks to identify ways to improve the overall effectiveness of cyber operations. Through this study, we intend to deeply explore the competitive and historical dynamics of Russian intelligence agencies, which are crucial to a comprehensive understanding of the broader landscape of cyber operations. In the following sections we add succinct, top-level descriptions of TTPs employed by the analysed APTs, for they serve as valuable tools in understanding their behaviour and modus operandi.

7. The Agencies Case Studies

The choice to focus on the GRU, SVR, and FSB agencies for the case study portion of our OCO study was driven by

several significant factors. Firstly, the context of the recent Russian-Ukrainian conflict at the centre of this paper, which has seen a marked increase in Russian cyber capabilities, makes these agencies particularly relevant. The GRU, SVR, and FSB have been protagonists in various cyber operations in this context. These agencies have distinct but complementary roles in intelligence and cyber operations. The GRU deals primarily with military intelligence, the SVR with foreign intelligence, and the FSB with internal security and counter-intelligence. By analysing the interactions between these agencies, we can gain greater insight into Russia's internal dynamics in cyber operations. Another critical aspect is the historic competition and disparities between these agencies. These internal differences offer a rich context for exploring how they influence coordination and effectiveness in cyber operations. Furthermore, our analysis focuses on the impact of this lack of coordination on the effectiveness of cyber operations. If these agencies fail to coordinate effectively, this could reduce the impact of their cyberattacks and increase the likelihood of detection. By examining interactions at operational and technical levels, our study seeks to identify ways to improve the overall effectiveness of cyber operations. Through this study, we intend to deeply explore the competitive and historical dynamics of Russian intelligence agencies, which are crucial to a comprehensive understanding of the broader landscape of cyber operations. In the following sections we add succinct, top-level descriptions of TTPs employed by the analysed APTs, for they serve as valuable tools in understanding their behaviour and modus operandi.

7.1. GRU

The Main Directorate of the General Staff of the Armed Forces of the Russian Federation, commonly called the GRU, is Russia's military intelligence agency. The GRU has been implicated in some of the best-known cyber operations, and the public profile of the units underscores a high operational pace. The GRU would also control several research institutes tasked with developing new malware. Over the years, researchers and analysts have noted an apparent willingness on the part of GRU computer units to conduct aggressive espionage operations, sometimes with questionable operational security and secrecy levels [33]. In particular, Unit 26165, to which, APTs, such as Fancy Bear and Sandworm, are linked, is one of the two Russian groups identified by the US government as responsible for hacking the DNC during the Clinton-Trump presidential campaign. Western governments and media have linked Unit 26165 to numerous offensive operations against public and

private sector targets in the United States and Europe [34]. Then there is Unit 74455, which is linked to some of Russia's most brazen and damaging cyberattacks. Unit 74455 was identified as responsible for the coordinated release of stolen emails and documents during the 2016 US presidential election [35]. Focusing primarily on systems penetration and intelligence gathering, Unit 74455 appears to have a significant offensive cyber capability, including developing NotPetya malware that hit multiple targets in Ukraine in June 2017, then spread globally and caused significant damage outside Ukraine [36]. Finally, there is Unit 54777, also known as the 72nd Special Service Center, which would be responsible for GRU psychological operations, including online disinformation campaigns [37].

(1) *Sandworm*: While Sandworm is not Kremlin's most prominent hacker group, it is the most visible one since the beginning of the war, and its track record of successful attacks with global impact, most notably the NotPetya malware and several attacks on Ukraine have made it a severe concern for the Computer Emergency Response Team of Ukraine (CERT-UA). In 2017, the group used Wiper NotPetya malware disguised as ransomware to take down hundreds of networks between Ukrainian government agencies, banks, hospitals, and airports, causing an estimated \$10 billion in global damage. By presenting destructive attacks as ransomware, Sandworm would be able to cover its tracks and make it more difficult for researchers to attribute the attacks to a state-sponsored group. Since the beginning of the war, Sandworm has relentlessly targeted Ukraine with various malware strains. Some were highly sophisticated, while others exploited known vulnerabilities that made them easier to detect and prevent from spreading. Researchers believe Sandworm experimented with malware strains to bypass Ukraine's best defences. Most of the attacks were neutralised in the early stages, and the second blackout researchers expected from Sandworm after targeting Ukraine's power supply in 2015 and 2016 never occurred [38]. In April 2022, Sandworm attempted to take down a large energy supplier in Ukraine using a new iteration of the 'Industroyer' malware dubbed 'Industroyer2' just for ICS systems, as well as a new version of the 'CaddyWiper' malware to destroy data of the organisations affected. According to reports, Industroyer2 has been customised to target high-voltage power substations and then use CaddyWiper and other malware for data wiping (e.g. OrcShred, Soloshred, and Awfulshred for Linux and Solaris systems) and then wipe any trace of the attack [39]. It is still unknown exactly how Sandworm compromised the energy supplier's environment or how it moved from the IT network, according to researchers at the computer company ESET, who worked

with CERT-UA to secure the network to the ICS environment. ESET strongly believes that Industroyer2 was created using the source code of Industroyer, exploited by Sandworm in 2016 to shut down power in Ukraine. According to CERT-UA and ESET, Sandworm planned to initiate the final phase of this attack by distributing the malware on April 8, 2022 on Azure servers and automated Windows workstations, Linux servers running OrcShred and AwwfulShred, high voltage power substations and active network equipment. CERT-UA points out, however, that the implementation of Sandworm's evil plan has so far been prevented, thanks to efficient operational detection and incident response planning. ESET also noted in a technical report on the malware used in the attack that 'Sandworm allegedly attempted to distribute Industroyer2 malware against high-voltage power substations in Ukraine'. ESET researchers further report that Industroyer2 is configurable and includes detailed hardcoded configuration, which requires it to be recompiled for each new target. ESET points out, however, that given that the Industroyer malware family has only been deployed twice, with a 5-year gap between each release, Sandworm operators still need to develop different versions. The malware sample shows functionality similar to Industroyer's IEC-104 module, primarily a protocol used in Europe and the Middle East for TCP communications within electrical systems. There are conflicting reports about the impact of this operation. While the full impact remains to be seen, this operation serves as a reminder of Russia's capabilities to cut off electricity in different parts of Ukraine and its readiness to employ them. This activity poses a higher risk to Ukraine's electricity transmission and distribution services [40]. Sandworm is also allegedly responsible for a new round of ransomware attacks hitting targets across Ukraine with the new variant of [the .NET](#) RansomBoggs ransomware. Also, ESET, in a series of tweets about ransomware attacks, claims to have informed CERT-UA of a variant of RansomBoggs that it spotted, as the ransomware targeted several local organisations. Reports indicate that the [exploited .NET](#) malware is new and distributed similarly to previous campaigns linked to GRU. The ransom note (SullivanDecryptsYourFiles[.].txt) shows the authors impersonating James P. Sullivan, one of the main characters in the Pixar film Monsters & Co. The executable file is also called Sullivan[.].exe. There are similarities to previous Sandworm attacks: a PowerShell script used to [distribute .NET](#) ransomware from the domain controller is nearly identical to the one seen last April during the Industroyer2 attack s against the energy sector, ESET researchers explain. The PowerShell script used, which CERT-UA dubbed 'PowerGap', was also used to distribute the 'CaddyWiper' malware alongside Industroyer2 using the 'ArguePatch' loader [41]. ESET

also says the operation resembles a ransomware campaign conducted in October 2022 that targeted Ukrainian and Polish logistics companies with the 'Prestige' variant. The ransomware's activity targeting Ukrainian organisations named RansomBoggs has not been directly observed. However, the PowerShell script used to distribute [the .NET](#) ransomware known as POWERGAP is tracked. This script can enumerate Group Policy Objects using the Active Directory service interface, in line with other recent activity involving NEARMISS, CADDYWIPER, and JUNKMAIL, all delivered via GPO. In particular, the activity that exploits these tools together with POWERGAP is attributed – at the time of writing – to APT28 too, which, like Sandworm, would be under the control of GRU [42].

(2) *Fancy Bear*: The cyber espionage activity of Fancy Bear, also known as APT28, Strontium, or Sofacy, has mainly targeted entities in the United States, Europe, and the countries of the former Soviet Union, including governments and armed forces, the media, and dissidents at the present Russian government. In recent years, Russia appears to have been using APT28 increasingly to conduct intelligence operations commensurate with broader strategic military doctrine. APT28 uses the same pattern to hit its victims: after compromising a victim organisation, APT28 steals sensitive data, which is then leaked for other political narratives aligned with Russian interests [43]. These have included the conflict in Syria, NATO-Ukraine relations, the European Union (EU) refugee and migrant crisis, and the 2016 US presidential election [44]. Since 2014, APT28's online activity has likely supported intelligence operations designed to influence the domestic politics of foreign nations. These operations have involved taking down and defacing websites, false flag operations using fake hacktivists, and data theft later publicly disclosed online. APT28 is also responsible for the attack on the DNC and other entities related to the 2016 US presidential election cycle. These breaches involved the theft of internal data, primarily emails, which were later strategically leaked through multiple forums and calculatedly propagated, almost certainly intended to further particular objectives of the Russian government [45]. In a report published on January 7, 2017, the US Office of the Director of National Intelligence (ODNI) [46] described this activity as an 'influence campaign'. This influence campaign – a combination of network compromises and subsequent data leaks – aligns closely with the Russian military's publicly stated intentions and capabilities. Influence operations, also often called information operations, have a long history of inclusion in the Russian strategic doctrine and have been intentionally developed, deployed, and modernised through the so-called

Gerasimov doctrine with the advent of the Internet. APT28 is believed to have played a significant role in the ongoing conflict in Ukraine, mainly through its cyber operations. The group has been linked to several cyberattacks against the Ukrainian government, including military targets and critical infrastructure, as well as disinformation campaigns designed to influence public opinion in the country [35]. APT28, as early as January 14, 2022, a month before the invasion, reported that the Google Threat Analysis Group (TAG) would have been the proponent of a phishing campaign focused on Ukraine. On March 16, 2022, CERT-UA issued an alert highlighting that UAC-0028, the name CERT-UA gave APT28, was phishing UkrNet accounts. On March 4, 2022, Microsoft reported that it also noticed that the government network in Vinnytsia, a city in west-central Ukraine, was compromised by APT28 through a vicious spear phishing campaign targeting Ukrainian military and Ukrainian government personnel in the region. On May 3, 2022, Fancy Bear was then observed targeting its victims with a new variant of infostealer malware, distributed via email attachments, while on May 6, 2022, CERT-UA issued a new alert on another campaign by 'APT, which allegedly sent malicious emails posing as the CERT-UA, containing an attachment in the form of a password protected RAR archive 'UkrScanner.rar' and inside the RAR file, a self-extracting archive (SFX) containing a malware called CredoMap. The data collected by the malware was exfiltrated via [HTTPPOST](#) requests to *.m.pipedream[.]nethostnames [47]. In particular, the CERT-UA warned that Sandworm, also linked to the Russian government, would collaborate with APT28 in these months of the conflict to target and actively exploit the vulnerability known as 'Follina' in Microsoft Windows Support Diagnostic Tool (MSDT) (CVE-2022-30190) in malspam attacks. According to CERT-UA, the malspam messages use subject lines, such as 'LIST of links to interactive maps' within a malicious Word document (e.g. LIST_of_links_in_interactive_maps[.]docx) and have already reached more than 500 recipients. The CERT-UA advisory reads that attackers continue to exploit the CVE-2022-30190 vulnerability and increasingly resort to emails from compromised government-domain emails. Ukrainian government experts have traced this activity to UAC-0113, a threat actor they say with medium confidence is associated with Sandworm. In reality, Mandiant keeps track of the activity reported publicly as UAC-0113 and believes, it is UNC3666, an undefined persistent threat which might be associated with APT28, with moderate confidence, and which serves explicitly to carry out everyday coordination activities between the two APTs for attacking the same targets. UNC3666 has likely targeted Ukrainian organisations as early as December 2021 [48].

7.2. SVR

The Foreign Intelligence Service (SVR) is Russia's principal civilian intelligence agency for foreign countries. Its task is to collect information using Human Intelligence (HUMINT), Signal Intelligence (SIGINT), and Cyber Intelligence (CYBINT) methods.¹ Most analysts conclude that SVR operates forcefully, emphasizing secrecy and detection avoidance [49]. Most cyber operations related to the SVR focus on intelligence gathering [50]. The SVR has high technical expertise, often trying to achieve and maintain persistence within compromised networks. Some computer analysts refer to SVR hackers as Cozy Bear or Turla [45].

(1) *Cozy Bear*: Cozy Bear, also known as APT 29, CozyDuke, the Dukes or PowerDukes, is a threat actor which has been active much earlier than the Russian-Ukrainian conflict, and is shown to have strong ties with the SVR since 2008. APT29 is also known to have been, together with APT28, involved in the US Democratic National Committee compromise in 2015. Following the 2016 US presidential election, APT29 was found responsible for spear-phishing campaigns targeting US-based governmental and non-governmental organisations (NGOs). The phishing emails were sent to defence, national security, international affairs, and law enforcement personnel. Some of the emails even pretended to originate from the Clinton Foundation to share election analysis. APT29 has continued to evolve and improve, showcasing new TTPs. Undoubtedly, APT29 has quite a diverse toolkit of custom-developed tools that continually improves as new information is published to the infosec community. This set of tools mainly focuses on gaining permanent access to the victim's machine through backdoors and harvesting information, files, credentials, etc. and their exfiltration. APT29 used a wide range of different programming languages to develop its malware, from pure Assembly (present in some components of the MiniDuke malware) to C++(CozyDuke) and from C#, [VisualBasic.NET](#) (HammerDuke and RegDuke) to Python (SeaDuke). The group's creativity goes even further, as they customise and try different technologies, infection vectors, infrastructures, and more [51]. In summary, APT29 represents a dangerous advanced persistent threat. The group is technically skilled and capable of adapting to the defences of its chosen targets. It often uses techniques and tools that have been identified in previous attacks. The 'fingerprints' of its attack activity are becoming well documented and the subject of considerable ongoing scrutiny [52]. Against the backdrop of the war in Ukraine, APT29 is exploiting a 'lesser-known' Windows feature called Credential Roaming following a successful phishing attack against a European diplomatic entity. The diplomacy-focused targeting is consistent

1——HUMINT (Human Intelligence) is intelligence obtained through human interaction, while SIGINT (Signal Intelligence) refers to intelligence gathered through the interception of signals. CYBINT (Cyber Intelligence) is a sub-category of intelligence involving collecting information from cyberspace for analysis and use in cyber security.

with Russian strategic priorities and APT29's historic targeting, as reported by Mandiant researcher Thibault Van Geluwe de Berlaere. APT29 is known for its intrusions aimed at gathering information in line with the strategic objectives of SVR [53]. Some of the collective's cyber activities are publicly monitored under the Nobelium moniker, a threat cluster responsible for widespread supply chain compromise through SolarWinds software in December 2020. Google said, it identified the use of Credential Roaming during the period APT29 was present within the victim's network in early 2022. Then, 'several LDAP queries with atypical properties' were executed against the Active Directory system. Introduced in Windows Server 2003 Service Pack 1 (SP1), Credential Roaming allows users to access their credentials securely on different workstations in a Windows domain. According to Microsoft, Credential Roaming stores user credentials in ms-PKI-DPAPIMasterKeys and ms-PKI-AccountCredentials in the user object. The latter is a multi-valued LDAP property containing a sizable binary object (BLOB) containing data and encrypted credentials. According to the TAG group, one of the LDAP attributes queried by APT29 concerned ms-PKI-Credential-Roaming-Tokens, which manages blob storage of encrypted user credential tokens for roaming [54].

(2) *Turla*: Turla, also known as Snake, Uroburos, Venomous Bear, or Waterbug, is the other group that, together with APT29, has links to the SVR, although, it is noteworthy that Microsoft places it within a cluster of known threats linked to FSB. Since at least 2007, this threat actor has allegedly been responsible for high-profile cyberattacks and espionage campaigns against government, military and diplomatic entities, research and defence organisations in Ukraine, and several NATO states. Turla is also known for its sophisticated and stealthy techniques, often using custom malware and advanced tools to infiltrate its targets' networks and remain undetected for long periods. Over the years, the collective has been involved in several high-profile cyber espionage campaigns, including campaigns in the United States, Europe, and the Middle East [55]. Some of the unique tools and malware used by Turla include the following:

Snake/Uroburos: A highly sophisticated root kit used for espionage and data exfiltration, capable of infecting both 32-bit and 64-bit systems. It is designed to run on infected systems for extended periods undetected.

KopiLuwak: A Javascript-based malware used in targeted attacks, which can perform various tasks, such as downloading and

executing additional payloads, communicating with specific command and control (C2) servers, and data exfiltration.

EpicTurla (also known as Wipbot or Tavdig): A modular backdoor that provides remote access to compromised systems and has been used in cyber-espionage campaigns since at least 2012 [56]. In a year of conflict, Turla was observed exploiting vulnerabilities in the systems of critical Ukrainian organisations and infrastructures with malware developed over a decade earlier to deliver reconnaissance tools and backdoors to specific targets in Ukraine. Mandiant, who has been monitoring APT's various operations since the beginning of the war, said that the malware used corresponds to a variant of a malware called ANDROMEDA (aka Gamarue), uploaded to VirusTotal back in 2013. Since the start of the Russian military invasion of Ukraine in February 2022, the collective was allegedly linked to a series of phishing and credential reconnaissance activities targeting various entities in the country. Among the incidents analysed by Mandiant, in one, an infected USB stick was used in a Ukrainian organisation as early as December 2021, leading, once inserted into the systems, to the distribution of ANDROMEDA on different hosts, thanks to the launch of a malicious link (.LNK) masquerading as a folder inside the USB drive [57]. The threat actor then repurposed one of the dormant domains of ANDROMEDA's defunct C2 infrastructure – re-registering the domain in January 2022 – to profile the victim by launching the KOPILUWAK dropper. Two days later, on September 8, 2022, the attack moved to its final stage with the execution of a .NET-based implant called QUIETCANARY (aka Tunnus), resulting in the exfiltration of all files created after January 1, 2021. Mandiant also allegedly identified a spyware application for Android masquerading as a 'Process Manager' service to stealthily steal sensitive information stored on infected devices. Interestingly, this app – has the package name 'com.remote.app' – establishes contact with a remote command and control server, 82.146.35[.]240, which has been identified as infrastructure belonging to Turla. When the application runs, a warning about the permissions granted to the application is displayed. Permissions include screen lock and unlock attempts, global device proxy settings, screen lock password expiration settings, storage encryption settings, and disabling cameras. Once the app has been activated, the malware runs in the background, abusing broad permissions to access device contacts, call logs, track device location, send messages, access external storage, take pictures, and record audio. The collected information is in JSON format and transmitted to the remote server. Also, unknown at this stage is the exact initial access vector used to distribute the spyware and the intended goals of the campaign. The rogue Android app

also attempts to download a legitimate application called Roz Dhan (meaning 'daily wealth' in Hindi), which has over 10 million downloads and allows users to earn cash rewards for completing surveys and questionnaires. In July 2022, however, TAG revealed that Turla would create another malicious Android app; this time, however, to support pro-Ukrainian hacktivists to launch Distributed Denial-of-Service (DDoS) attacks against Russian sites. This activity by Turla dovetails with what has been written so far to support the group's casualty profiling efforts coinciding with the Russo-Ukrainian war and SVR interests, helping the agency gather information of interest to the Russian government [58].

7.3. FSB

The Federal Security Service, or FSB, is Russia's principal internal security agency, responsible for internal security and counterintelligence. The FSB's tasks are protecting Russia from foreign cyber operations and monitoring domestic cybercriminal groups, a mission undertaken jointly with Department K of the Ministry of Internal Affairs [59]. In recent years, the FSB has expanded its remit to include foreign intelligence gathering and OCOs. Today's state-sponsored hacker groups linked to the FSB are Callisto, EnergeticBear, Gamaredon, TeamSpy, Dragonfly, Havex, CrouchingYeti, and Koala. SBU intelligence analysts say that the FSB has two primary centres overseeing information security and cyber operations. The first is the 16th Center, which houses most of the FSB's intelligence capabilities. The second is the 18th Center for information security, which oversees operations within national borders, but also conducts operations abroad. Like the GRU, the FSB oversees dedicated training and research institutes, which directly support the agency's offensive activities. Most of the operations appear to be reconnaissance or clandestine surveillance [60]. In 2021, Ukrainian intelligence released information and recordings about Crimean-based 18th FSB Center officers as part of the Gamaredon hacker group. Media reports indicate that this FSB unit is capable of developing advanced malware, and modifying known malware to imitate other APTs to hide their activities. Here we limit our analysis to the two main APTs linked to FSB: Callisto and Gamaredon.

(1) *Callisto*: Callisto has been an APT focused on cyber espionage at least since 2015. Over the years, this group has targeted various organisations, including government institutions and military officials in Eastern Europe and the South Caucasus. The APT uses spear-phishing campaigns and social engineering tactics to inject

malware into its targets. The group has also been observed to use remote access trojans (RATs) and credential-stealing malware to exfiltrate sensitive information from their victims. Callisto (aka COLDRIVER) is suspected to be a Russian APT which – although not publicly linked with any Russian intelligence service – has, in past operations, been shown to have objectives which align closely with the strategic interests of the FSB. Callisto mainly focuses on specific Western countries, namely, the United States and Eastern European countries [61]. During the conflict in Ukraine, the group master-minded several phishing campaigns aimed at stealing credentials, targeting areas of military and strategic research, such as NATO entities and defence entities based in Ukraine, as well as NGOs and think tanks. Additional targets include former intelligence officials, experts on Russian affairs, and Russian citizens abroad. While the SBU, the Security Service of Ukraine, has publicly associated Callisto with the Gamaredon group – which we discuss in the next section – through a set of hacks attributed to the FSB and essentially focusing on operations in Ukraine since the start of the Russian invasion in February 2022, other security companies do not support this link [62]. In particular, the IT security company SEKOIA.IO has conducted numerous technical investigations, not finding any overlap between the activities of Callisto and Gamaredon, nor any coordination or cooperation activity between the two APTs, indicating a lack of intra-agency coordination. They instead suggest that these are two groups operating on different targets and purposes. Based on what SEKOIA.IO investigated, domains aligned with Callisto’s past activities. Further investigations resulted in a more extensive infrastructure of more than 80 domains, including domain typosquatting activities. Since many of these domains were already known and the IP address resolution was already attributed to Callisto’s activities, SEKOIA.IO only associated these domains with Callisto with high confidence. In campaigns observed in the past, Callisto sent malicious PDF attachments to their victims. The first page of the PDF simulated an error in the PDF renderer engine, prompting the victim to open a link that led to a malicious web page. This web page was tasked with collecting the victim’s credentials using EvilGinX. Placing the phishing link in a PDF, rather than in the body of the email, prevents the link from being parsed by email gateways and is an effective tactic to remain undetected from an attacker’s perspective. SEKOIA.IO conducted open-source research on typosquatted domains to identify targets. Six private companies based in the United States and Eastern Europe, and four NGOs were identified, all involved in supporting Ukraine. Most of the targeted private organisations engage in activities related to military equipment, military logistics, or humanitarian support for Ukraine, including a US

company that supplies humanitarian logistics and possibly tactical equipment to Kyiv. Other industries include information technology and computer security. SEKOIA.IO notes that all the targets identified so far through the investigation, namely, the industrial and military entities affected and the individuals involved in Russian affairs, are in line with Calisto's interests. Callisto also targets support which is not directly related to Ukraine. Among Calisto's malicious domains discovered, three have caught the attention of analysts, namely, `mvd-redir[.]ru` and `dns-mvd[.]ru` (high confidence), which are most likely a typosquatting of the Russian Interior Ministry, and `lk-nalog-gov[.]ru` (with low confidence), the Russian Federal Tax Service. Because Callisto has been observed to target Russian individuals overseas, SEKOIA.IO finds it plausible that Callisto also engages in domestic surveillance activities. Another, less plausible, hypothesis would be a false flag manoeuvre to raise doubts about the attribution of the infrastructure. SEKOIA.IO found another potential victim that matches Callisto's known targeting. The domains `sangrail-share[.]com` and `sangrail-ltd[.]com` are typosquatting Sangrail Inc., a private security company, registered in the United Kingdom on July 31, 2019, by Ian Walter Baharie. That name was also used to register AC21, a British private intelligence firm focused on African politics [63]. Interestingly, this name appeared in a 17-year-old data leak that exposed a list of several MI6 officers on cryptome.org, a website dedicated to information leaks. That observation matches Microsoft's assessment of Callisto targeting former intelligence officers. It should be assessed that this kind of intrusion is aimed at a targeted collection of information contributing to the Russian efforts to interrupt the supply chain of military reinforcements for Kyiv. Nonetheless, SEKOIA.IO estimates that Callisto contributes to intelligence gathering for Russian intelligence on identified evidence related to war crimes or international justice proceedings, likely to anticipate and build a counter-narrative about future allegations. Among Callisto's targets, there would also be NGOs and European and international institutions, evidence that this type of activity could enter the sphere of competence of the SVR and would indicate competitive activity between this agency and the FSB.

(2) *Gamaredon*: Gamaredon's activity as an APT has been observed since 2013. It is believed to have ties with FSB, specifically Unit 71330. Although Gamaredon and Dragonfly are two separate APTs, both may be related to Unit 71330. While Gamaredon mainly focuses on cyber espionage and intelligence gathering, Dragonfly (also known as EnergeticBear or Crouching Yeti) is reportedly notorious for sophisticated and multi-stage attacks aimed at compromising

industrial control systems (ICS) and control systems of supervision and data acquisition (SCADA). Furthermore, while both groups may share TTPs, such as the use of spear-phishing emails as an initial attack vector, there is no direct evidence to suggest that they are related or operate jointly. Gamaredon uses a variety of techniques and tools to compromise its targets, including, as already mentioned, spear-phishing emails with malicious attachments, social engineering attacks, and exploitation of known software vulnerabilities (n-days). Some of the malware and tools used by the Gamaredon group include Pteranodon, Jupyter, and PowerShell-based tools [64]. In more detail, Gamaredon uses PowerShell scripts to automate various tasks, such as malware distribution, privilege escalation, and data exfiltration. Since the Russian invasion of Ukraine, the group remains one of the critical cyber threats to Ukrainian cyberspace. Gamaredon would operate from Sevastopol in Russian-occupied Crimea, acting on orders from the FSB's Center for Information Security in Moscow. The group began operations in June 2013, just months before Russia annexed the Crimean Peninsula from Ukraine. In its recent information-gathering campaigns against Ukraine, Gamaredon used malware written in PowerShell, known as GammaLoad and GammaSteel. These data exfiltration tools manage to capture files of specific extensions, steal user credentials, and take screenshots of the victim's computer. These two pieces of malware are not new and were previously used by Gamaredon to target Ukraine's government and security services. Hackers use phishing emails to gain initial access to the victim's network. These emails contain malicious LNK files distributed in RAR archives. Only users with Ukrainian IP addresses can open these files. Hackers send phishing emails from domains associated with legitimate organisations, such as the Security Service of Ukraine, and the names of the malicious files included are usually associated with the war in Ukraine. Gamaredon's recent activity is characterised by the multi-stage distribution of malware payloads used to maintain persistence. These payloads represent similar variants of the same malware, each designed to behave the same way as the others. According to CERT-UA, Gamaredon's TTPs would have evolved during the war, improving its tactics and retraining the malware variants used to go undetected. CERT-UA said [41] that Gamaredon is responsible for the most significant cyberattacks in Ukraine (even higher than those carried out by Sandworm), recording more than 70 incidents related to the group in 2022. Gamaredon also attacks allies of Ukraine. Latvia confirmed a phishing attack on its defence ministry in late January 2022, linking it to the group. Ukrainian cybersecurity officials described their attacks as intrusive and daring, and said the group's primary

purpose is to conduct targeted cyber intelligence operations [54]. Case study analysis of OCOs conducted by the Russian GRU, SVR, and FSB agencies highlights a complexity and sophistication that transcends the execution of conventional cyberattacks. In the context of the Russian-Ukrainian conflict, however, it emerged how the APTs linked to these agencies exploited their distinctive skills to implement operations, highlighting a level of internal coordination, which, precisely because of the inevitable tensions and divergences, significantly influenced the effectiveness and the extent of their actions in cyberspace. The case study investigation not only enriches our understanding of the operational TTPs peculiar to the Russian cyber offensive but also highlights how the lack of coordination can limit the overall impact of operations in the digital domain. Due to this lack of uniform coordination, the ability to operate highlights a strategic dimension that can surprisingly work against Russian offensive capabilities in cyberspace.

8. Conclusions

This evolving, descriptive paper scrutinises the intricate coordination within intelligence agencies, with a particular emphasis on the Russian landscape. The study is methodically structured around two principal RQs that guide the exploration of this complex domain. RQ1 seeks to unravel: ‘To what degree is integration between technical and operational levels achieved within intelligence agencies responsible for executing offensive government policies in cyberspace?’ This inquiry casts light on the multifaceted nature of coordinating cyber operations that engage numerous state-endorsed APTs managed by various intelligence units. The coordination challenges identified encompass a spectrum of technical dilemmas, including system compatibility, software intricacies, network issues, and timing delays. Additionally, it examines strategic complications, such as the intersection and potential conflict of objectives and methodologies among different agencies, which could escalate into issues of territorial and power disputes. RQ2 examines: ‘What elements hinder the integration between technical and operational levels in intelligence agencies tasked with enacting government defensive strategies in cyberspace?’ This query delves into the impediments to effective inter-agency cooperation, highlighting factors like varying organisational cultures and operational dynamics. Issues such as disparities in trust-building, leadership styles, decision-making processes, and management of uncertainties are explored, as these can lead to misalignments in objectives and misunderstandings. The paper also addresses the critical ‘principal-agent’ dynamic, wherein intelligence agencies (agents) have

greater informational access than decision-makers (principals), leading to potential reluctances in information sharing and negatively impacting strategic decision-making and intelligence operations. The research uncovers the profound rivalry among Russian intelligence agencies, notably FSB, SVR, and GRU, marked by their overlapping roles and internal competitions. This environment, coupled with the necessity for cohesive coordination in cyber operations, unveils a host of technical, strategic, and human-centric challenges [65]. While this study has focused on specific organisational, cultural, and operational factors impeding coordination between intelligence agencies, it is important to acknowledge that there may be additional elements at play. These could include geopolitical considerations, budgetary constraints, and technological disparities. The rapidly evolving nature of cyber threats and technologies may also contribute to coordination challenges, as agencies may struggle to keep pace with new developments and adapt their strategies accordingly. Furthermore, the broader political landscape and national security priorities can significantly influence inter-agency dynamics. Changes in government administration, shifts in foreign policy, or emerging global threats may alter the balance of power and responsibilities among intelligence agencies, potentially exacerbating existing coordination issues or creating new ones. As a work in progress, this research paves the way for a multitude of future inquiries. These prospects span various methodologies and themes within the cyber intelligence field, encompassing the study of organisational behaviours in intelligence agencies, the analysis of collaborative mechanisms between different agencies, and the exploration of strategies to effectively navigate the complex dynamics inherent in state-sponsored cyber operations. In conclusion, while the coordination of APTs across multiple intelligence agencies holds significant potential to enhance the impact of cyber operations, it is entangled with a series of formidable challenges. Addressing these challenges necessitates an all-encompassing grasp of the nuances in cyber operations, an acknowledgment of the cultural and operational variances among agencies, and adept management of the 'principal-agent' dynamic. Only through a comprehensive approach to these factors can intelligence entities fully harness the capabilities of coordinated cyber operations [2].

References

- [1] D. Štrucl, "Russian aggression on Ukraine: Cyber operations and the influence of cyberspace on modern warfare," *Contemporary Military Challenges (Sodobni Vojas'ki Izzivi)*, vol. 24, pp. 103–123, 2022, doi: [10.33179/bsv.99.svi.11.cmc.24.2.6](https://doi.org/10.33179/bsv.99.svi.11.cmc.24.2.6).

- [2] M.S. Weiss, "Russian Military Intelligence: Background and Issues for Congress," CRS Report R46616, 2021. [Online]. Available: <https://crsreports.congress.gov/product/pdf/R/R46616>. [Accessed: Nov. 9, 2023].
- [3] A. Smith, "Public-Private Partnerships and Collective Cyber Defence," in Proceedings of the IEEE International Conference on Cyber Security and Cybercrime (ICCCS), M. Thompson, R. Johnson, Ed. New York: IEEE, 2022, pp. 75–85, doi: [10.23919/CyCon55549.2022.9810912](https://doi.org/10.23919/CyCon55549.2022.9810912).
- [4] F. Ebinger, S. Veit, N. Fromm, "The partisan–professional dichotomy revisited: Politicisation and decision-making of senior civil servants," Public Administration, vol. 97, no. 4, pp. 861–876, 2019, doi: [10.1111/padm.12613](https://doi.org/10.1111/padm.12613).
- [5] M. Alderighi, C. Feder, "Institutional design, political competition and spillovers," *Regional Science and Urban Economics*, 2020, doi: [10.1016/j.regsciurbeco.2019.103505](https://doi.org/10.1016/j.regsciurbeco.2019.103505).
- [6] J. Moses, "Political rivalry and conflict in Putin's Russia," *Europe-Asia Studies*, vol. 69, pp. 961–988, 2017, doi: [10.1080/09668136.2017.1364700](https://doi.org/10.1080/09668136.2017.1364700).
- [7] S. Taillat, F. Douzet, "Collective security and strategic instability in the digital domain," *Contemporary Security Policy*, vol. 40, pp. 362–367, 2019, doi: [10.1080/13523260.2019.1602693](https://doi.org/10.1080/13523260.2019.1602693).
- [8] L.M. Maguire, "Managing the hidden costs of coordination," *Communications of the ACM*, vol. 63, pp. 90–96, 2020, doi: [10.1145/3379989](https://doi.org/10.1145/3379989).
- [9] J.M. Ostrow, "Conflict-management in Russia's federal institutions," *Post-Soviet Affairs*, vol. 18, pp. 49–70, 2002, doi: [10.1080/1060586X.2002.10641513](https://doi.org/10.1080/1060586X.2002.10641513).
- [10] J.J. McNeil, *Maturing international cooperation to address the cyber space attack attribution problem*, PhD Dissertation, Norfolk, VA: Old Dominion University, Norfolk, VA, 2010.
- [11] J.L. Hernandez-Ardieta, J. Tapiador, G. Suarez-Tangil, "Information sharing models for cooperative cyber defence," in *5th International conference on cyber conflict (CYCON 2013)*, June 2013.
- [12] E.V.D. Heuvel, G. Klein Baltink, "Coordination and Cooperation in Cyber Network Defense: The Dutch Efforts to Prevent and Respond," in *Best Practices in Computer Network Defense: Incident Detection and Response*, R. Badger, P. Thompson, Eds. Berlin: Springer, 2014, pp. 35–50, doi: [10.3233/978-1-61499-372-8-118](https://doi.org/10.3233/978-1-61499-372-8-118).
- [13] T. Liebetrau, "Organizing cyber capability across military and intelligence entities: Collaboration, separation, or centralization," *Policy Design and Practice*, vol. 6, no. 2, pp. 131–145, 2023, doi: [10.1080/25741292.2022.2127551](https://doi.org/10.1080/25741292.2022.2127551).
- [14] A. Ahmad, J. Webb, K.C. Desouza, J. Boorman, "Strategically motivated advanced persistent threat: Definition, process, tactics and a disinformation model of counterattack," *Computers & Security*, vol. 86, pp. 402–418, 2019, doi: [10.1016/j.cose.2019.07.001](https://doi.org/10.1016/j.cose.2019.07.001).
- [15] *Defending Ukraine: Early lessons from the cyber war*. [Online]. Available: <https://blogs.microsoft.com/on-the-issues/2022/06/22/defending-ukraine-early-lessons-from-the-cyber-war/>, 2022. [Accessed: May. 15, 2023].

- [16] M. Khaleefa, M. Abdulah, "Concept and difficulties of advanced persistent threat," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 16, no. 6, pp. 29–35, 2016. [Online]. Available: <https://www.semanticscholar.org/paper/Concept-and-difficulties-of-advanced-persistent-Khaleefa-Abdulah/c0e8fb235c9bdfba5a066fdbba4ae5a660dc0fa8>. [Accessed: Nov, 23 2023].
- [17] S. J. Shackelford, M. Sulmeyer, A.N. Craig, B. Buchanan, B. Micic, "From Russia with love: Understanding the Russian cyber threat to U.S. critical infrastructure and what to do about it," *Conflict Studies: Terrorism eJournal*, 2017.
- [18] R. Simonson, J.R. Keebler, M. Lessmiller, T. Richards, J. Lee, "Cyber Security Teamwork: A Review of Current Practices and Suggested Improvements," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, M. Matthews, S. Landry, A. Bisantz, Eds. Thousand Oaks, CA: SAGE Publications, 2020, vol. 64, pp. 451–455, doi: [10.1177/1071181320641101](https://doi.org/10.1177/1071181320641101).
- [19] D.V. Gioe, "Cyber operations and useful fools: the approach of Russian hybrid intelligence," *Intelligence and National Security*, vol. 33, pp. 954–973, 2018, doi: [10.1080/02684527.2018.1479345](https://doi.org/10.1080/02684527.2018.1479345).
- [20] J. Cheravitch, B. Lilly, *Russia's Cyber Limitations in Personnel Recruitment and Innovation: Their Potential Impact on Future Operations and How NATO and Its Members Can Respond*. Santa Monica, CA: RAND Corporation, 2020.
- [21] F.T. Sheldon, G. Peterson, A. Krings, R. Abercrombie, A. Mili, *Proceedings of the 5th Annual Workshop on Cybersecurity and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies*. New York, NY: ACM Press, 2009.
- [22] G. Bonnet, C. Tessier, "Coordination despite constrained communications: A satellite constellation case," in *Proceedings of the 4th International Conference on Space Mission Challenges for Information Technology (SMC-IT 2008)*, Pasadena, CA, USA, 2008, pp. 91–98. [Online]. Available: <https://www.semanticscholar.org/paper/Coordination-despite-constrained-communications-%3A-a-Bonnet-Tessier/89110ff5f7e68700bd069f197c2662d1292de0fe>. [Accessed: Feb. 20, 2023].
- [23] J.-H. Eom, "Roles and Responsibilities of Cyber Intelligence for Cyber Operations in Cyberspace," *Computer Science and Engineering*, vol. 19, no. 3, 2014, pp. 45–56.
- [24] E. Iasiello, "What is the role of cyber operations in information warfare?" *Journal of Strategic Security*, vol.14, no. 4, pp. 72–86, 2021, doi: [10.5038/1944-0472.14.4.1931](https://doi.org/10.5038/1944-0472.14.4.1931).
- [25] J. Rollins, C. Wilson, "Terrorist Capabilities for Cyberattack: Overview and Policy Issues," Congressional Research Service, Washington, DC, USA, 2007. [Online]. Available: <https://www.semanticscholar.org/paper/Terrorist-Capabilities-for-Cyberattack%3A-Overview-Rollins-Wilson/4cb91489579c09e0b191f579b4605748e2376604> [Accessed: Feb. 21, 2023].
- [26] A. Samojlova, "Social engineering methods," *Scientific Development Trends and Education*, 2019, doi: [10.18411/j-11-2019-48](https://doi.org/10.18411/j-11-2019-48).
- [27] R. Egnell, "Civil-military coordination for operational effectiveness: Towards a measured approach," *Small Wars & Insurgencies*, vol. 24, no. 2, pp. 237–256, 2013, doi: [10.1080/09592318.2013.778017](https://doi.org/10.1080/09592318.2013.778017).

- [28] C. Clough, "Quid pro quo: The challenges of international strategic intelligence cooperation," *International Journal of Intelligence and Counter Intelligence*, vol. 17, no. 4, pp. 601–613, 2004, doi: [10.1080/08850600490446736](https://doi.org/10.1080/08850600490446736).
- [29] T.H. Hammond, "Why is the intelligence community so difficult to redesign? Smart practices, conflicting goals, and the creation of purpose-based organizations," *Governance*, vol. 20, pp. 401–422, 2007, doi: [10.1111/j.1468-0491.2007.00364.x](https://doi.org/10.1111/j.1468-0491.2007.00364.x).
- [30] K. Kralovanszky. (2021). *Certain connections between cyber operations, artificial intelligence and operational domains*. Hadtudoma'nyiSzemle Hadmu've'szet. [Online]. Available: <https://orcid.org/0000-0002-5560-3525> [Accessed: Jul. 19, 2023].
- [31] G. Hofstede, *Culture's consequences: International differences in work-related values*, vol. 5. Los Angeles, CA: Sage, 1984.
- [32] E. Meyer, *The Culture Map: Breaking Through the Invisible Boundaries of Global Business*. New York, NY, USA: Public Affairs, 2014.
- [33] K. Giles, "'Information Troops' – A Russian Cyber Command?" in *Proceedings of the 3rd International Conference on Cyber Conflict (CyCon 2011)*, Tallinn, Estonia, 2011, pp. 1–16.
- [34] I. Ciosek, "Aggravating Uncertainty – Russian Information Warfare in the West," *Torun International Studies*, vol. 13, no. 1, 2020, pp. 75–88, doi: [10.12775/TIS.2020.005](https://doi.org/10.12775/TIS.2020.005).
- [35] A.J. Dawson, M. Innes, "How Russia's internet research agency built its disinformation campaign," *The Political Quarterly*, vol. 90, no. 2, pp. 245–256, 2019, doi: [10.1111/1467-923X.12690](https://doi.org/10.1111/1467-923X.12690).
- [36] S. Goel, "Cyberwarfare: Connecting the dots in cyberintelligence," *Communications of the ACM*, vol. 54, pp. 132–140, 2011, doi: [10.1145/1978542.1978569](https://doi.org/10.1145/1978542.1978569).
- [37] K. Pynno'niemi, *Information-psychological warfare in Russian security strategy*. London and New York: Routledge, 2019, doi: [10.4324/9781351181242-21](https://doi.org/10.4324/9781351181242-21).
- [38] F.J. Egloff and M. Smeets, "Sandworm: A new era of cyberwar and the hunt for the Kremlin's most dangerous hackers," *Journal of Cyber Policy*, vol. 5, no. 2, pp. 326–327, 2020, doi: [10.1080/23738871.2020.1808032](https://doi.org/10.1080/23738871.2020.1808032).
- [39] A. Greenberg. (Apr. 12, 2022). "Russia's sandworm hackers attempted a third blackout in Ukraine," *Wired* [Online]. Available: <https://www.wired.com/story/sandworm-industry-attack-ukraine/> [Accessed: Dec. 5, 2023].
- [40] A. Scroxtion, "Sandworm rolls out industroyer2 malware against Ukraine," *Computer Weekly.com*, Apr. 12, 2022. [Online]. Available: <https://www.computerweekly.com/news/252515855/Sandworm-rolls-out-Industroyer2-malware-against-Ukraine> [Accessed: Jan. 12, 2023].
- [41] D. Antoniuk. (Nov. 29, 2022). "Sandworm hacking group linked to new ransomware deployed in Ukraine," *The Record* [Online]. Available: <https://therecord.media/sandworm-hacking-group-linked-to-new-ransomware-deployed-in-ukraine> [Accessed: Jan. 13, 2023].
- [42] R.M.A. Molina, S. Torabi, K. Saredidine, E. Bou-Harb, N. Bouguila, C.M. Assi, "On ransomware family attribution using pre-attack paranoia activities," *IEEE Transactions on Network and Service Management*, vol. 19, pp. 19–36, 2022, doi: [10.1109/TNSM.2021.3112056](https://doi.org/10.1109/TNSM.2021.3112056).

- [43] A. Lemay, J. Calvet, F. Menet, J.M. Fernandez, "Survey of publicly available reports on advanced persistent threat actors," *Computers & Security*, vol. 72, pp. 26–59, 2018, doi: [10.1016/j.cose.2017.08.005](https://doi.org/10.1016/j.cose.2017.08.005).
- [44] D.L. Linvill, B.C. Boatwright, W.J. Grant, P.L. Warren, "'The Russians are hacking my brain!' investigating Russia's internet research agency twitter tactics during the 2016 United States presidential campaign," *Computers in Human Behavior*, vol. 99, pp. 292–300, 2019, doi: [10.1016/j.chb.2019.05.027](https://doi.org/10.1016/j.chb.2019.05.027).
- [45] H. Mwili, T. Dargahi, A. Dehghantanha, K.-K.R. Choo, "Analysis and Triage of Advanced Hacking Groups Targeting Western Countries' Critical National Infrastructure: APT 28, RED October, and Regin," in *Handbook of Big Data and IoT Security*, A. Dehghantanha, R.M. Parizi, K.-K.R. Choo, Eds. Cham: Springer, 2019, pp. 245–266, doi: [10.1007/978-3-030-00024-0_12](https://doi.org/10.1007/978-3-030-00024-0_12).
- [46] S. Slick, "The Role of the Director of National Intelligence as Head of the Intelligence Community," Foreign Policy Research Institute, September 2019. [Online]. Available at: <https://www.fpri.org/article/2019/09/the-role-of-the-director-of-national-intelligence-as-head-of-the-intelligence-community> [Accessed: Feb. 4, 2023].
- [47] J. Burt, "Russia's Apt28 targets Ukraine government with bogus Windows updates," *The Register*, May 2, 2023. [Online]. Available at: https://www.theregister.com/2023/05/02/russia_apt28_ukraine_phishing/ [Accessed: May 10, 2023].
- [48] V. Kumar, C. Shah, "Countering Follina Attack (CVE-2022-30190) with Trellix Network Security Platform's Advanced Detection Features," Trellix, Jul. 19, 2022. [Online]. Available: <https://www.trellix.com/en-us/security-news.html>. Accessed: [Accessed May 10, 2023].
- [49] R.F. Staar, C.A. Tacosa, "Russia's security services," *Mediterranean Quarterly*, vol. 15, pp. 39–57, 2004, doi: [10.1215/10474552-15-1-39](https://doi.org/10.1215/10474552-15-1-39).
- [50] I. Thornton-Trump, "Russia: The Cyber Global Protagonist," EDPACS: The EDP Audit, Control and Security Newsletter, vol. 65, no. 2, pp. 19–26, 2022, doi: [10.1080/07366981.2022.2041226](https://doi.org/10.1080/07366981.2022.2041226).
- [51] G. Brogi, V. Viet Triem Tong, "Terminaptor: Highlighting advanced persistent threats through information flow tracking," in: *8th IFIP international conference on new technologies, mobility and security (NTMS)*. pp. 1–5, 2016, doi: [10.1109/NTMS.2016.7792480](https://doi.org/10.1109/NTMS.2016.7792480).
- [52] E.M. Hutchins, M.J. Cloppert, R.M. Amin, "Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains," Lockheed Martin Corporation, 2010. [Online]. Available at: <https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/LM-White-Paper-Intel-Driven-Defense.pdf> [Accessed: Nov. 17, 2022].
- [53] R. Waters. (Nov. 10, 2022). *Apt29 using Windows credential roaming bug to target diplomats. Mandiant finds Apt29 increasingly targeting NATO and its allies in 2022*. [Online]. Available: <https://www.cybercareers.blog/2022/11/apt29-using-windows-credential-roaming-bug-to-target-diplomats/> [Accessed: Dec. 20, 2023].
- [54] R. Lakshmanan. (Nov. 9, 2022). *Apt29 exploited a Windows feature to compromise European diplomatic entity network*. [Online]. Available: <https://thehackernews.com/2022/11/apt29-exploited-windows-feature-to.html> [Accessed: Nov. 10, 2022].

- [55] D. Pereira. (Jun. 7, 2023). *The origin story of the Aptturla, the hunt for' the snake' malware, and current steps for prevention*. [Online]. Available: <https://www.oodaloop.com/archive/2023/06/07/the-origin-story-the-fsbs-turla-the-hunt-for-the-snake-malware-and-current-steps-for-prevention/>. [Accessed: Jul. 10, 2023].
- [56] Securelist by Kaspersky. (Aug. 7, 2014). *The epicturla operation*. [Online]. Available: <https://securelist.com/the-epic-turla-operation/65545/>. [Accessed: Dec. 10, 2023].
- [57] L. Gyongyosi. (Jan. 9, 2023). *Turlause sold malware infrastructure to attack Ukrainian institutions: Andromeda USB spreading malware used for data exfiltration*. [Online]. Available: <https://heimdalsecurity.com/blog/turla-uses-old-malware-attack-ukrainians/>. [Accessed: Nov. 20, 2023].
- [58] B. Leonard. (Jul. 19, 2022). *Continued cyber activity in Eastern Europe observed by threat analysis group*. [Online]. Available: <https://blog.google/threat-analysis-group/continued-cyber-activity-in-eastern-europe-observed-by-tag/>. [Accessed: Dec. 10, 2022].
- [59] L. Turkaeva, "Federal security service in the national security system," 2020, doi: [10.20310/2587-9340-2020-4-15-399-406](https://doi.org/10.20310/2587-9340-2020-4-15-399-406).
- [60] J. Kose, "Cyber Warfare: An Era of Nation-State Actors and Global Corporate Espionage," Semantic Scholar. [Online]. Available: <https://www.semanticscholar.org/paper/Cyber-Warfare%3A-An-Era-of-Nation-State-Actors-and-Kose/d10e2841df8e35c85830d69e54fc262c4e01ebe9>. [Accessed: Jan. 18, 2023].
- [61] K.S.R. Rani, B.C. Soundarya, H.L. Gururaj, V. Janhavi, "Comprehensive analysis of various cyberattacks," in: *IEEE Mysore Sub-section International Conference (MysuruCon)*, 2021, pp. 255–262.
- [62] F. Aimé, M.A. Togun, "Calisto Shows Interest in Entities Involved in Ukraine War Support," *Cyber Threat Intelligence Bulletin*, Dec. 5, 2022.
- [63] I.Group®. (Sep. 19, 2022). *Russia-nexusuac-0113 emulating telecommunication providers in Ukraine*. [Online]. Available: <https://www.recordedfuture.com/russia-nexus-uac-0113-emulating-telecommunication-providers-in-ukraine>. [Accessed: April. 7, 2023].
- [64] G. Tiepolo. (Feb. 14, 2023). *Russian apt'gamaredon' exploits hoax shell to target Ukrainian organizations*. [Online]. Available: <https://mrtiepolo.medium.com/russian-apt-gamaredon-exploits-hoaxshell-to-target-ukrainian-organizations-173427d4339b>. [Accessed: May 26, 2023].
- [65] B. Lilly, J. Cheravitch, "The Past, Present, and Future of Russia's Cyber Strategy and Forces," in *Proceedings of the 12th International Conference on Cyber Conflict (CyCon)*, T. Minárik, R. Jakschis, L. Lindström, Eds. Tallinn: NATO CCD COE Publications, 2020, pp. 189–203, doi: [10.23919/CyCon49761.2020.9131723](https://doi.org/10.23919/CyCon49761.2020.9131723).

Post-Truth and Information Warfare in their Technological Context

Ignas Kalpokas | Department of Public Communication, Vytautas Magnus University, Lithuania, Institute of Information Systems and Digital Economy, SGH Warsaw School of Economics, Poland | ORCID: 0000-0003-1110-5185

Abstract

As citizens are faced with an overabundance of information, their reliance on intuitive sorting strategies and platform-enabled content selection and delivery increases correspondingly. Under such circumstances, political action tends to be based on haphazard encounters with opinion-congruent content than on anything else, giving rise to so-called post-truth condition and, in turn, opening up conditions for manipulating such information encounters as part of information warfare operations. In particular, this novel environment necessitates a rethinking of informational agency, locating it within interactions between humans and technological artefacts, whereby humans as generators of data and algorithms as tools that structure the information domain based on such data co-construct political and social spaces. The impact of digital technologies is further amplified by the advent of synthetic (Artificial Intelligence-generated) media, which is foreseen to bring about epistemic confusion, that is, increasing inability to separate between reality and fiction. Under such conditions, and in any situations of actual or perceived crisis and tension, audiences are inclined to rely on narratives as coping strategies, which is where information warfare operations come to the fore. Either capitalising on the existing fertile ground or having manufactured a condition of crisis and distrust, such operations are geared towards hijacking audience cognitive processes with narratives that suit their perpetrators.

Received: 15.11.2023

Accepted: 20.06.2024

Published: 27.07.2024

Cite this article as:

I. Kalpokas "Post-truth and information warfare in their technological context," ACIG, vol. 3, no. 2, 2024, pp. 99–121. DOI: 10.60097/ACIG/190407

Corresponding author:

Ignas Kalpokas,
Department of Public
Communication, Vytautas
Magnus University,
Lithuania, and Institute
of Information Systems
and Digital Economy,
SGH Warsaw School of
Economics, Poland; E-mail:
ignas.kalpokus@vdu.lt
 0000-0003-1110-5185

Copyright:

Some rights reserved:
Publisher NASK



Keywords

generative AI, post-truth, information warfare, epistemic confusion, trust

1. Introduction

The war in Ukraine has proven to be a testing ground for new and emerging military technologies, such as drones. However, besides the kinetic battlefield, warfare operations have also been taken in the information domain. Notably, such operations have been making use of technological developments to a no lesser extent than their kinetic counterparts. Hence, the aim of this article is to explore transformations in digital communication that have enabled a qualitatively new breed of information warfare. In order to do so, this article is built on a conceptual review of the existing trends and developments with the aim of developing a conceptual framework for explaining the interaction between post-truth, information warfare, and Army Intelligence (AI)-based technologies. In order to do so, key ideas and recent developments regarding post-truth, changes in information environment, and the advent of AI-based synthetic media are identified and their connections elucidated. The identified transformations are subsequently connected to the key features of information warfare campaigns.

Of course, discussions of manipulation, disinformation, and the receding importance of veracity have been the focal point for communication studies for quite some time, often focusing on post-truth. As such, post-truth is best seen as collusion between audiences, technology companies, and political actors, whereby audiences derive both satisfaction and information benefits (such as quick navigation in an oversaturated information environment) but in exchange open themselves to manipulation [1]. Meanwhile, information warfare is broadly understood as a deliberate effort by state and non-state actors to shape the strategic environment within a particular public sphere or across multiple public spheres in a way that suits the perpetrator's interests [2]. In essence, the aim is to affect the thought processes of general populations or political elites (or both) so that decisions are made using the frames, preconceptions, and habitual associations implanted by and commensurable with the interests of the perpetrator [2]. Crucially, information warfare leaves no room for a strict war/peace dichotomy characteristic of western thinking – instead, it is always on, taking place in the background, even though it tends to be amplified in situations of crisis or conflict when low-level nudging is deemed by the perpetrator to be no longer sufficient [2].

Post-truth and information warfare can be seen as cousin concepts that share similar premises but differ in terms of intentionality. Post-truth refers to a general transformation of the information environment and an ensuing reconfiguration of the relationship between veracity and political beliefs and action. However, deviations from truth generally happen organically, as a result to the transformations of the information ecosystem. Information warfare, meanwhile, refers to the deliberate and strategic manipulation of the information environment that makes use of, among other things, post-truth tendencies to guide audiences towards predetermined patterns or thinking. It is, therefore, crucial to understand the specificities of both post-truth and information warfare as well of the information ecosystem upon which they are jointly premised.

In order to conceptualise contemporary information warfare and contextualise it within the broader post-truth condition, this article proceeds in four parts. Firstly, the emergence of post-truth as a result of the changing information environment is overviewed. This is followed by a more in-depth analysis of technological transformations, namely in the second part, the de-centring of humans in communication processes and, in the third part, the likely emergence of epistemic confusion due to proliferation of synthetic media. Finally, these strands are taken together in a discussion of information warfare strategies.

2. Post-Truth and the Changing Information Environment

One of the core changes at the heart of the formation of today's information environment has been the shift in emphasis from a supposed 'information age' towards a 'post-truth era'. In general terms, post-truth is understood as a tendency by audiences to opt for opinion-congruence and ease of access/cognition instead of veracity as the main criteria for information selection. This has been associated with changing patterns of information supply (ever-growing amount of content, replacement of professionally prepared and curated content with user-generated content, and algorithmic content governance) as well as societal factors, such as politicians and other actors making use of such conditions in ways that are contributing to societal polarisation. Post-truth has largely been brought about by the ever-growing interdependence between humans and digital technologies. Indeed, while previously the Internet was itself seen as a 'liberation technology', enabling networked individuals to self-organise in a struggle for democracy and freedom [3], currently the attention has shifted to the opposite end

of the spectrum, namely manipulation, disinformation, and information warfare. These are seen to be not only features of domestic political competition (a domain usually associated with post-truth) but also international competition and even hybrid warfare strategies. In the case of Russia's war against Ukraine, the same can also be seen as an addition to conventional warfare practices. While warfare, propaganda, and attempts to 'win hearts and minds' have traditionally gone together, the interplay between warfare and post-truth leads to more pervasive, all-encompassing, and interactive practices in the management of audience cognitive processes.

A key concern in contemporary information and communication studies is that 'we are witnessing historical changes in the process of production of knowledge, characterised by high velocity and dizzying excess, as well as the development of new forms of digitally derived knowledge' [4, p. 26]. While one might take issue with the epochal scale of such assertions, it is, nevertheless, clearly the case that not only the amount of available content has overtaken the capacity to pay attention (which, in fact, is not new) but also the speed with which content changes and new items are added goes beyond the abilities to keep track and make sense. The preceding has been further exacerbated by the disaggregation of news supply in the context of social media: instead of competing as collective offerings (a newspaper, a news broadcast, etc.), news and other media content currently compete as standalone de-contextualised items, resulting in increased competition and hampering of content selection [5]. As this information environment is also devoid of traditional gatekeepers and open to an almost unlimited flow of user-generated content, sense-making capacity is only further overwhelmed [6], meaning that '[t]he challenge of communication overload is that each message can be heard – as the carrier of a distinct meaning – yet it cannot be attended to, since the time required for doing so is lacking' and leading to the need for individuals to 'drastically select from the environment' so that attention can only be paid to what seems to be noteworthy [7, pp. 112, 113]. All precedencies make automated content governance a valuable function performed by digital platforms [8], thus underscoring the importance of choices of and by algorithms.

Clearly, digital content, including news and other information, is 'ubiquitous, pervasive, and constantly around us', ultimately driving individuals to expect news to find them instead of seeking information intentionally [9, p. 106]. In other cases, people may become so overwhelmed and anxious about the ever-increasing stream of news that they begin avoiding them altogether, further deepening

their dependence on piecemeal haphazard encounters [10]. The preceding directly implies that attention is both a scarce and pivotal resource in the present media environment [11]. It thus should not come as a surprise that audiences have become spoilt for choice: as opinion-congruent content is always available, regardless of the level to which it corresponds to verifiable facts, selective exposure to information increasingly becomes the norm [6]. Moreover, such selective sorting is further strengthened by the online platforms themselves, whereby content-to-be-liked is algorithmically selected and displayed to any given user. Consequently, the current transformations of the public sphere have the tendency to result in fragmentation into opinion-congruent bubbles [12]. Such need for opinion-congruence can also be abused by way of manufacturing false unanimity through automated accounts and other forms of manipulation [13]. It is also notable that citizens are by far not mere passive recipients of digital information flows and the algorithmic logics inherent therein but are also active in the generation and spread of such content, thus at least partly taking agency into their own hands – for better or worse, often engaging in what has been called ‘participatory propaganda’ [14].

Attention capture is further implicated with the algorithmic processes of information delivery, particularly insofar as social media platforms are concerned. The latter processes are predicated upon personalised targeting of content so that individuals are permanently offered that they are bound to like and pay attention to, leading towards ‘the growing personalization of constructed realities and the subsequent individualization effects’ [15, p. 254]. Hence, as a direct consequence of the overabundance of information and competition over attention, citizens’ worldviews become further strengthened and entrenched through imaginary confirmation of their pre-existing beliefs. Crucially, then, in the digital environment described above, it transpires that the quality of information is far less important in driving political participation than the feeling of being informed, meaning that those driven by deficient information are just as likely to make their voices and opinions heard and actively push for opinion-commensurable political decisions as those who possess verifiably factual knowledge, thereby leading to further proliferation of a-factual points of view¹ and their inclusion in the political agenda [16], thus contributing to post-truth politics.

Nevertheless, one needs to resist the dominant temptation in literature on post-truth towards ‘clear-cut distinctions between the esteemed objective realm of facts, science, and reason and the dangerous subjective realm of emotions, ideology, and irrationality’

1———‘A-factual’ is used here as an inclusive term to accommodate both the more organic straying aside from truth concerns (‘post-truth’) and intentional disinformation.

[17, p. 787]. Simultaneously, the willingness in some recent revisionist literature to dismiss the idea of post-truth as merely a ‘moral panic’ [18] is unproductive as well, because it simply recasts the narrative in progressivist terms and, therefore, fails to engage with the critical potential of the idea of post-truth. In particular, it is important to understand that the condition, typically referred to as ‘post-truth’, is a consequence of the digital information ecosystem, rather than determined by the inner deficiencies of the individuals that happen to be following and supporting a-factual narratives. Hence, such individuals must not be marginalised and looked down upon (which, again, is common in the literature on post-truth) but, instead, the factors that have led them to their particular beliefs have to be investigated. It is far from uncommon for such factors to include information warfare operations. The latter, however, must not be taken as a universal category either: instead, just like warfare in general, information warfare makes use of technological transformations and developments, which today involve significantly transformed interrelationships between humans and digital technologies.

3. Digital Communication Environment: Moving Beyond Human-Centricity

As already intuited in the previous part, accounting for changes in the communication environment are crucial in order to understand the socio-political processes in today’s societies. Broadly, the communication environment is understood here as the sum total of technological and other means for sending and receiving information (in terms of both private interactions and matters of public concern) available to a particular society at a given time and combined with the predominant use practices on behalf of the audiences. With an ever-increasing role of digital technologies and various AI-enabled tools and algorithmic governance mechanisms, today’s communication environment has not only grown in terms of complexity but is also putting in question some of the often taken-for-granted assumptions about human-centricity in communication. Of course, such human-centricity largely remains intuitive: after all, intentionality and the capacity to generate and understand meaning within specific contexts are all central to communicative interactions. Simultaneously, though, AI tools now have significant sway over the public arena by way of shaping the information received by individuals (e.g. content selection and moderation), generating part of the content consumed by individuals, and even acting as communication partners, such as in the case of voice assistants [19]. The crucial questions, however, revolve around

the depth and kind of such technological participation. It must be stressed, however, that the thrust behind this section is diagnostic: instead of celebrating or criticising the tendencies described above, the aim is to contribute to the understanding of the latter.

Human-technology interrelatedness is manifested in the structure of today's public arena, best understood in general terms as 'interconnected communicative spaces' [19, p. 165]. More precisely, should one attempt to break down the public, with Hasebring, Merten, and Behre, into constellations of actors, frames of relevance, and communicative practices, it becomes clear that AI-enabled technological artefacts participate in all of them [20]. They participate in publics alongside humans as both assistants and obstructors (bots could be an example of the latter), shape relevance by subterraneously structuring information supply, and take part in content generation and other practices that set frames for interaction. Other models paint an even more fragmented picture by focusing on communicative formations that are 'variously private and public, personal and topical, small and large, transient and persistent', being 'connected both horizontally and vertically by shared participants and information flows' [21, p. 79]. Moreover, it is not just the internal dynamics and user practices of such formations that determine their fate: instead, a crucial role is played by 'platform affordances, commercial and institutional interests, technological foundations, and regulatory frameworks' [21, p. 79], clearly implying a constant flux that is simultaneously shaped internally and externally. Here, again, the triple role of digital artefacts – as moderators of online encounters with content (e.g. platform affordances), interlocutors (bots, conversational agents, etc.), and content generators – comes to the fore. It thus should come as no surprise that in many ways, algorithms can function as partners in communication, for better or worse [22].

Notably, one could reasonably assert the emergence of the new normal in terms of 'construction of reality with and through digital media and infrastructures' [23, p. 147]. The preceding is, of course, a very general assertion, covering the broad societal transformations that are taking shape vis-à-vis digital technologies. A crucial issue at hand, though, is whether one can meaningfully discuss human-AI partnership in communication without the advent of Artificial General Intelligence. One way of tackling the problem could be reframing the question from one concerning AI to that of artificial communication; hence, it is not imitation of human intelligence (which remains elusive) but reproduction of communication skills that matters [22]. In this way, a fundamentally interactive model emerges: one of

enmeshment between human-generated data, machine learning processes, and communicative practices, even without the need to emulate human intelligence beyond the narrow domain of communication. Given the human–digital interdependence as the key premise of post-truth, such further enmeshment can be seen as deepening the replacement of veracity with outcomes of digital content flows as the benchmark for political and societal processes.

The preceding precludes one-sided assertions of loss of human agency and emergence of ‘algorithm dependency’ [24], pointing instead towards mutual dependence. When engaging with AI-enabled tools, the crux of the matter ‘is not that a human would interact with the material vis-à-vis a machine, but with systems that generate their communication based on a variety of human digital traces’ [23, p. 146]. The process is interactive: an AI tool would reflect the perspective of human actors as an aggregate but always with a twist – a perspective that enables such tools to interact with humans not by simply parroting them but also by producing an outcome that strikes a balance between recognisability and surprise; such outcomes, in turn, become a source of human interaction and learning, thus informing future interactive outcomes [22]. Once again, interrelatedness and enmeshment are evident. The environment thus produced ‘follows users’ choices, then processes and multiplies them, and then re-presents them in a form that requires new choices’ [22, p. 64]. In other words, AI-enabled tools react to and around humans (AI passivity, human activity) but do so in ways that externally structure the conditions for human behaviours and responses (human passivity, AI activity). Once again, post-truth is here best seen as an interactive condition.

Still, however, one might posit that there is a crucial difference, due to the agency of digital artefacts being, at best, conditioned by humans or even illusory. Nevertheless, it must be stressed that the centrality and independence of human agency has also come under intense questioning in recent years. Notably, today’s increasingly digital-first life means that the nature of the human self, let alone its supposedly autonomous qualities, is increasingly distributed among multiple data doubles – ‘de-corporealised’ virtual individuals residing within technology [25, p. 159]. The ensuing ‘human–data assemblages’ are in a constant state of flux ‘as humans move through their everyday worlds, coming to contact with things such as mobile and wearable devices, online software, apps and sensor-embedded environments’ [26, p. 466], conditioning them and being conditioned in return. It thus becomes evident that subjectivity and agency cannot be understood as autonomous

qualities describable in binary terms (as either present or absent) but, instead, best seen as in-between states [27]. The exceptionality of the human subject is, consequently, put to question. Consequently, one must acknowledge that ‘not only humans but also non-humans [...] have agentic and performative capacities’ [28, p. 380], resulting in shared abilities that are ‘more-than-human’ [29]. It indeed transpires that instead of the rational-autonomous ideal, ‘[w]e are relational beings, defined by the capacity to affect and be affected’, constantly ‘flowing in a web of relations with human and non-human others’ [30, pp. 45, 47]. Consequently, agency would thus be found in an ‘interplay of human capabilities and the capacities of more or less smart machines’ [31, p. 3]. One should, therefore, talk not of an increase or diminution of agency on either side of the human–AI encounter but, instead, of complex and dynamic networks of agency, with truth (or, rather, what counts as the latter) becoming immanent to such interactions.

The above view is also supported by neuroscientific research that reveals an autonomous unified self to be merely an illusory unity brought together out of diverse elements: multiple interacting neural networks, social interactions, and artefacts encountered at any given moment [32]. Hence, even the workings of human brain are best seen as an endless exercise in improvisation at the interplay between the external world and the memories of past thoughts and experiences instead of some manifestation of ‘a hidden inner world of knowledge, beliefs, and motives’ [33, p. 9]. Seen in this way, the relationship of being shaped by any encounter at hand and shaping the environment back through interpretation and reaction to such encounters (instead of linear autonomous human progress) is, simply, a natural feature and not a technologically conditioned one. Consequently, humans are merely entities constantly scrambling for meaning, undergoing a constant process of re-invention, rather than self-sufficient actors exerting power and dominance over their environments. Again, moving into the technological domain, then, the aim should be to move ‘beyond the competition narrative about humans and machines’ [34, p. 42] and avoid simplistic dualisms that merely obfuscate the complexities of contemporary societies characterised by mediatization [23, p. 147]. Overall, the goal should be to overcome binary thinking, instead aiming for an approach that would posit interactivity between humans and their environment as the default condition of communicative interactions. Under such conditions, another binary – between fact and fiction – is destabilised as well.

Overall, then, while the growing role of AI and algorithmic tools in communication has become a truism, it is time to move further by

positing horizontal interrelationship and enmeshment between humans and digital artefacts. On the one hand, this is due to the growing role and capacities of digital artefacts as structuring actors, interlocutors, and content co-generators; on the other hand, this is also consequent to autonomous human agency, traditionally taken for granted, emerging as, at best, an overstretch. In combination, a new, enmeshment- and interaction-focused, take-on communication and sense-making (on both individual and collective levels) is necessary. Likewise, the same pertains to any obstructions and complications in the flow of information or the poisoning of such flows through injection of disinformation. Seen from this perspective, one should focus less on alleged loss of some human mastery (the typical focus of mainstream approaches to post-truth), but, instead, on co-originating forms of content indistinguishability, including those that allow information warfare operations to hide in their midst.

4. Synthetic Media and Emerging Epistemic Confusion

In order to fully appreciate the role of technological developments in the emergence of post-truth and the creation of conditions for contemporary information warfare strategies, one must also consider the effects of artificial content generation. Indeed, the rise in prominence and growing adoption of generative AI has been one of the defining features of the past several years. While beneficial uses of this technology, including in communication, are plentiful, there are, nevertheless, clear security implications that need to be taken into account. Here, particular attention is typically paid to the potential use of AI generators to produce disinformation and deceive outrightly. However, instead of focusing on singular disinformation campaigns (which, it must be admitted, may pose significant threats but are, nevertheless, likely to remain isolated occurrences), more attention should be paid to underlying background effects caused by the very presence (and increasing prevalence) of AI-generated content. In broad terms, such effects could be described as epistemic confusion.

The subject matter here is synthetic media, namely ‘audio-visual media which has been partly or fully generated/modified by technology’ [35, p. 2]. Some key features to note here include democratisation of content creation (as easy-to-use interfaces enable users to leverage AI to generate content they would otherwise be unable to produce), increased speed and efficiency with which content is created, and the capacity to generate realistic yet fake depictions

of individuals and events. Crucially, regardless of the intention with which such synthetic content is generated, the mere fact of its omnipresence would likely lead to a diminishing of trust as individuals become increasingly unsure of whether the authenticity of the content they encounter can be reasonably established; moreover, particularly in situations when individuals are simply casually scrolling through available content, they may lack both time and attention to check and verify [36, 37]. Notably, it is not only de-contextualised pieces of information shared on social media that have to be treated with suspicion – entire websites masquerading as news sources filled with AI-generated text, featuring nonsensical content or outright falsehoods are already not uncommon [38]. In some cases, the aims behind resorting to synthetic content can be noble, such as attempts to counter disinformation by building AI tools that generate rebuttals – from social media posts to, again, entire websites staffed by fake journalists [39]. The downside, nevertheless, is that all of this only further stretches the cognitive load of individuals as they attempt to navigate online information spaces. Even in cases when synthetic content is not outrightly harmful and had not been created with a nefarious aim (including satire or parody), it can still have negative effects simply by lingering at the back of one's mind: not least, the very possibility that something has been AI-generated can reduce trust even in genuine information [35].

Crucially, the epistemic confusion induced by synthetic media is further strengthened by the dominant modes of content distribution. For example, algorithmic content governance on social media is by no means news-centric; moreover, such platforms tend to supply users with de-contextualised and entertainment-focused pieces of content, which precludes the formation of an effective representation of the societal issues at hand [40]. Users need to put in deliberate effort by intentionally seeking news content for this aptitude to be picked up by the algorithm. In other words, to paraphrase Gil de Zúñiga et al., news may still 'find me' [9], but only to the extent that I have made a head start. Nevertheless, as news are enmeshed with entertainment and other types of content for which the threshold of acceptable AI augmentation (or complete generation) is significantly lower, context differentiation and epistemic trust in news could well recede. Contexts themselves are likely to blur as the need to compete in a non-news-centric environment could also push informational content creators to turn to synthetic media to simply retain some relevance. All of this creates favourable conditions for actors engaged in information warfare operations by making cognitive overload and news cynicism among target audiences easier to achieve anything, including causes and atrocities of

war, can be caught within (or deliberately pushed towards) this spiral of indeterminacy.

Even when content is not shared but, instead, generated for personal use, such as consulting large language models (ChatGPT, Bard, etc.), increasing reliance on technologically mediated access to the world might lead not only to diminution of agency but also to the threat of uncritically accepting the output thus generated, despite its occasional propensity to falsehood [37], let alone data poisoning, adversarial attacks, and other hostile attempts by outside actors to negatively affect the output of such tools [41, 42]. Even short of hostile actions from outside, deterioration of outputs could happen due to ‘data inbreeding’, that is, AI models being trained on AI-generated data, which might happen either accidentally or by design as the proportion of online synthetic content continues to grow [43]. As user experience of the flaws and dangers of such models grows, their trust in any form of available knowledge and the possibility of distinguishing between truth and falsehood would likely suffer.

In addition to already familiar problems, extended reality environments may introduce a completely new set of threats, such as the potential to create false memories and introduce overlays that are difficult to distinguish from objective reality – both highly problematic in light of the accumulating neuroscientific knowledge that human perception of reality is based on predictive processing of the human brain that provides, effectively, best guesses and approximations of reality, rather than detached objective knowledge [44]. Hence, extended reality can be seen as having the potential to cause ‘disruption of deliberation between people due to the breakdown of a common reality’ [44, p. 11], thus further contributing to epistemic confusion. Indeed, the loss of shared touchpoints and increasing sufficiency of digital life could lead to the breakdown of even the fragmented and intermeshed public spaces that currently still allow some interconnections among citizens.

Certainly, efforts are underway to ease the cognitive load and, therefore, reduce epistemic confusion, with watermarking attracting the most attention. Still, while the thrust towards watermarking and otherwise identifying AI-generated content (both in terms of industry standards and regulatory frameworks, such as the European Union’s AI Act) is commendable, such measures can be undone through the use of specialised software (such as watermarks being either removed or made less prominent for human or machine detection, e.g. by the adding noise); moreover, for

content that mixes different media (e.g. text, audio, video, and images all being used in a single post on, say, TikTok), separation of authentic and fake is going to be even more difficult [36]. No less importantly, watermarks are only effective when AI-generated (or modified) content is the exception and not the norm: if the majority of content is synthetic, it is unlikely that watermarks would retain signifying value – that would merely become part of the fabric of everyday life, no longer drawing individuals' attention. Even more problematically, reliance on watermarks as a verification tool may induce a false sense of security: unwatermarked fake content (either with watermarks removed or produced using in-house tools, particularly by state and state-backed threat actors that have sufficient resources and sophistication) would automatically earn extra credibility. Not least, though, verification techniques can be abused through reverse watermarking, that is, adding fake watermarks that imitate common standards onto authentic content in an attempt to discredit it. Indeed, watermark manipulation can well open up a new front of information warfare.

The latter point captures a crucial aspect of epistemic confusion that is likely to follow the widespread adoption of synthetic media: as everything and anything can potentially be fake, the authenticity of anything can be put to doubt [36]. In fact, this does not even have to involve manipulation of authentic content so that it looks fake (such as adding a misleading watermark): in fact, mere accusation that an item has been digitally manipulated or AI-generated is sufficient to reduce trust and commitment [35, 45]. Falsely labelling content as AI-generated can happen both unintentionally (when people are over-vigilant, particularly vis-à-vis content they do not agree with) and deliberately (as a convenient way to dismiss content that goes against one's interests). Notably, the effects of such misleading accusations of fakery transpire to be stable over time and, crucially, have a greater effect on those who care about the particular topic at hand, perhaps because of their higher internal motivation to be adequately informed [45]. Hence, the threshold for deliberate manipulation of audience opinions is only further lowered.

5. Post-Truth, Information Warfare, and the Abuse of Coping Techniques

Conditions, identified here as post-truth, are particularly conducive to information warfare, particularly when taken in combination with the recent technology-driven changes in the information environment. In particular, the increasingly indeterminate

role of veracity and the changing contours of information agency extend the ambit of information warfare. In particular, this is due to the potential for abuse of coping techniques that, while not necessarily consciously employed by individuals, do nevertheless have significant leeway on how we understand our environment. Hence, avenues are opened if not for full conviction, then for further sowing of confusion among target audiences.

In order to better understand the coping mechanisms under conditions of uncertainty and how they could lead to the proliferation of information warfare operations, one needs to focus on the importance of a narrative. Crucially, it must be noted that people need a narrative because it ‘provides explanations’, that is, ‘describes the past, justifies the present, and presents a vision of the future’ [46, p. 120]. However, such a narrative is not always present at hand, particularly in times of rapid change or in crisis situations, which could be a natural disaster, an epidemic, a war, or anything of the like. In addition, as shown above, epistemic confusion can also be caused, or at least exacerbated, by technological factors, either independently or when they are strategically amplified. Under such conditions, pre-existing narratives no longer function and new explanations of the world are necessary. Since fact-based narratives may be slower to emerge (due to changing conditions and the need to establish the facts themselves beforehand), it is often difficult to fill the gap with verifiable information and an opportunity is created for alternative accounts to emerge, particularly if they produce a more satisfying (easier to comprehend and opinion-congruent) effect [47]. Indeed, what matters is the provision of meaning to an otherwise seemingly disorienting and disconcerting reality [48], even if that means falling for disinformation and succumbing to information warfare operations. After all, individuals expect from a narrative that it provides actionable insights, regardless of its veracity [49]. Moreover, it must also be noted that even fact-incongruent narratives have the capacity to ‘connect people, give meaning to experienced disparities and corruption in society’ [17, p. 785], particularly when they connect to grievances that often do have a factual basis and that had not yet been adequately explained or addressed.

Even more fundamentally, there are indications that the need and capacity to establish patterns even when none exist or when there is incomplete data to foresee their existence is hardwired through evolution [50], thus even further strengthening the need for explanatory or pseudo-explanatory narratives [48] and increasing the benefits to be accrued should such narratives be strategically

placed, centric for example, as a means of information warfare. Crucially, such behaviour helps individuals overcome the perceived randomness and complexity that otherwise typically characterise the world by providing order and predictability, however imaginary [50] and regardless of the broader political and societal implications. Of course, this could easily be dismissed as a normatively flawed coping strategy [50], and a lazy one for that matter, one merely concerned with 'simple recipes for explaining complex realities' [51, p 85]. It is, nevertheless an efficient solution in situations when information is either too scarce [52] or, on the contrary, too abundant [53], again at least from an individualist subjective perspective.

The preceding is particularly topical with regards to information warfare campaigns, carried out by both state and non-state actors, the aim of which is often to sow confusion and disorientation, for example, through hoaxes, fake news, and even plain scaremongering to subsequently make use of the ensuing collective action problems. Indeed, the first step of the process tends to be erosion of trust, both horizontally among citizens and vertically between citizens and their state/government, thereby creating fertile conditions for further hostile actions to be carried out [54], including nudging individuals towards specific narratives strategically placed to respond to pre-sown confusion. Once a spiral of distrust is set in motion by a threat actor, societies effectively enter a self-destruct mode, as the ensuing disorientation and polarisation makes it impossible (or at least very difficult) for citizens to formulate common interests and engage in achievement of any goals [2]. In fact, it might suffice to simply flood a selected public with competing contradictory opinions in order to diminish trust in any claims [55], very much in line with epistemic confusion described in the previous section. Moreover, it is important to note that trust increases openness to one's own vulnerability (thereby diminishing the need to rush for explanations and confusion-reducing narratives) and to other people's opinions (thus, potentially, also to corrections of one's own misperceptions); conversely, erosion of trust increases the likelihood of both falling for strategically placed narratives and becoming entrenched in one's own point of view [56].

Resorting to social media platforms for information warfare also enables threat actors to induce seemingly spontaneous audience reactions in response to messaging and to do so relatively simply, quickly, and at low cost. No less importantly, once successfully injected into the target audience, the manipulative message is propagated by citizens themselves (those who have become convinced

of its veracity), thereby further intensifying its spread [54]. Hence, herding target audiences into information silos or hijacking the existing filter bubbles constitutes a key strategic aim [55]. Threat actors then step in to resolve any uncertainty (including that of one's own making) and thereby both induce and respond to audience's need for comprehending any given situation and knowing how to act in the changing environment, particularly as such publics resort to unverified information should other or more quickly actionable options be unavailable [57]. Meanwhile, fact-based interventions to counter post-truth and/or information warfare operations may not only be at a disadvantage but could also derail the entire veracity-focused narrative by making it more complex and disorienting, thereby paradoxically increasing the demand for clear-cut, albeit less factual, stories that seemingly put all things in order [46]. What the preceding indicates, then, is that '[t]ruth, as in a fact or piece of information, has no intrinsic value'; rather, it can be claimed that '[i]t is up to the narrative to create that value' [46, p. 124]. Hence, the core variable for success, especially in the political domain, 'is not evidence (i.e. facts) but meaning' [58, p. 73]. Consequently, there are ample opportunities for the spread of conspiracy theories [58] or deliberative disinformation efforts, such as information warfare operations.

Sometimes neither full internalisation of a coherent narrative nor sowing confusion but affecting the perception of one's standing in the society might be the aim. In this case, establishment of immediate associations (positive or negative) attached to certain political and societal actors would likely end up affecting citizen modes of participation as well as perceptions of government policies, ethnic or other groups, general sense of societal development, etc. [56]. The preceding often relies on generating a sense of marginalisation. Here, it is crucial to keep in mind that one of the drivers that motivate resorting to factually false narratives is powerlessness and lack of control, either actual or perceived [50]. This typically involves groups that are societally underprivileged and lack a subjectively convincing possibility for emancipation or groups that had previously been privileged but have since been displaced or are being pushed aside by new, more progressive, groups, meaning that their concerns are also likely to be ignored or dismissed. Of course, in some cases such underprivileged status might be grounded in objective reality, but perceptions of such state of affairs could equally be manufactured as well. Likewise, groups that are disproportionately affected by ongoing crises (economic, health, military, etc.) can be more susceptible to disinformation and attempts to mislead. Strategically manufactured narratives would then

be aimed at providing perceived solutions by offering a sense of belonging to a community of those allegedly in the know, thereby bringing about a sense of subjective empowerment [51]. The latter, then, also brings inter-group dynamics into the mix as individuals are inclined to think that they and their group are firmly rooted in reality, making biases and false assumptions particularly difficult to spot (if they pertain to in-group views) and fostering polarisation by way of externalising the blame to non-like-minded others [59]. Hence, falling for fake news, disinformation, and information warfare operations tends to be understood by individuals and their peer groups as something that 'others' do, leading to the perception that others are vulnerable; the preceding then leads to another dichotomy: the self/we as seemingly rational and critically minded, and of the other as, allegedly, less intellectually gifted [60]. Such contrast can also lead to a false sense of security, whereby the intellectually superior self is seen as resilient by default and in a lesser need to care about the premises of one's own thinking.

It must also be kept in mind that proliferation of false narratives has been made possible by the general drive towards datafication, characteristic of contemporary societies: as populations are rendered fundamentally knowable by way of ubiquitous data collection, their pain points, biases, and preconceptions become relatively easy to identify [61]. The preceding has also significantly transformed the way in which political and opinion leadership is commonly understood: from being at the forefront of audience thought processes to following and voicing them [61]. Audience expectations are also not immune to such transformations as audiences simply expect to be satisfied, rather than challenged. Notably, there is an important international dimension here as well since crisis situations, particularly global ones, also imply the need for a sense of direction, community values, and shared identities, all of which are typical targets of information warfare [56]. Likewise, a key aim on either side of information warfare operations is to create positive habitual perceptions and a sense of shared concerns/values with one's own side in the minds of strategically targeted global audiences while fostering a sense of dissociation with one's adversary, either on a global or regional level [56]. Again, it is not only full convincing but sowing distrust and doubt within an adversary's support network that could be seen as a strategic goal.

Crucially, though, it is important to keep in mind that the effects of information warfare operations tend to be cumulative, meaning that they only become evident over time, once disintegration of a state's informational public (and, consequently, public order) or

global support network becomes manifest – that is, when the harm has already been done [56] and achievement of strategic goals, both domestically and abroad, has been impeded [55]. In this way, protection from such operations becomes particularly problematic. While much of the response has thus far concentrated on proactive defence measures, such as media and information literacy, their effectiveness has thus far faced only very limited empirical testing and lacks reliability due to the absence of a control group. Therefore, the offence should be seen as continuing to maintain an advantage within the domain of information warfare.

6. Conclusions

Overall, it must be noted that the changes in contemporary information environment, particularly overabundance of content and its algorithmic management, has led to a transformation of the role of veracity. In many ways, what is taken as truth and, therefore, as actionable, has become contingent upon attention management strategies employed by individuals, group dynamics, and, most importantly, data-based automated matching of individuals and content that the former are predisposed to like. To this effect, humans must be seen as sharing information agency with an increasing array of digital tools. Such structural conditions are also favourable to information warfare operations that can exploit the new patterns of content dissemination and consumption in order to inject strategically carved narratives into the minds of selected audiences. Moreover, the rapid spread of synthetic media is beginning to initiate yet another change – the emergence of epistemic confusion, whereby everything and anything could potentially be manipulated. Under such conditions, demand for seemingly stable and coherent explanatory narratives can be seen as a coping strategy, with information warfare operations being geared towards offering such alleged solutions. Moreover, deliberate erosion of trust (with the consequent retreat from mainstream information and increased need for explanatory narratives) often happens to be the first stage of information warfare, creating the conditions to nudge target audiences towards pre-crafted narratives – which is all the easier within the present technological context. Overall, then, it transpires that technological change and the ensuing transformations in the information domain have created a new strategic environment in which states targeted by information warfare operations are constantly on the back foot, with limited solutions to ameliorate this situation.

Of course, similar tools and techniques can be used not only to proliferate disinformation but also by strategic communications and

other counter-disinformation agents. However, in terms of epistemic confusion, it is by no means clear yet, what the end societal effect would be (reduction of potentially harmful beliefs vs further increased epistemic confusion). It is a matter for future research to establish the balance between, for example, mere uncertainty-inducing epistemic confusion versus disinformation-weakening epistemic confusion.

Funding

This work did not receive any specific funding.

References

- [1] I. Kalpokas, *A political theory of post-truth*. London: Palgrave Macmillan, 2018.
- [2] I. Kalpokas, J. Kalpokiene, "Synthetic media and information warfare: Assessing Potential Threats," in *The Russian Federation in the global knowledge warfare*, H. Mölder et al. (eds.), Cham: Springer Nature, 2021, pp. 33–50.
- [3] L. Diamond, "Liberation technology," *Journal of Democracy*, vol. 21, no. 3, pp. 69–83, 2010, doi: [10.1353/jod.0.0190](https://doi.org/10.1353/jod.0.0190).
- [4] P. Dahlgren, "Media, knowledge and trust: The deepening epistemic crisis of democracy," *Javnost – The Public*, vol. 25, no. 1, pp. 20–27, 2018, doi: [10.1080/13183222.2018.1418819](https://doi.org/10.1080/13183222.2018.1418819).
- [5] K. Munger, "All the news that's fit to click: The economics of Clickbait Media," *Political Communication*, vol. 37, no. 3, pp. 376–397, 2020, doi: [10.1080/10584609.2019.1687626](https://doi.org/10.1080/10584609.2019.1687626).
- [6] R. McDermott, "Psychological underpinnings of post-truth in political beliefs," *PS: Political Science & Politics*, vol. 52, no. 2, pp. 218–222, 2019, doi: [10.1017/S104909651800207X](https://doi.org/10.1017/S104909651800207X).
- [7] N. Couldry, A. Hepp, *The mediated construction of reality*. Cambridge: Polity, 2017.
- [8] S. Vaidhyanathan, *Anti-social media: How Facebook disconnects us and undermines democracy*. Oxford: Oxford University Press, 2018.
- [9] H. Gil de Zúñiga, B. Weeks, A. Ardèvol-Abreu, "Effects of the news-finds-me perception in communication: Social media use implications for news seeking and learning about politics," *Journal of Computer-Mediated Communication*, vol. 22, no. 3, pp. 102–123, 2017, doi: [10.1111/jcc4.12185](https://doi.org/10.1111/jcc4.12185).
- [10] B. Toff, R.K. Nielsen, "How news feels: Anticipated anxiety as a factor in news avoidance and a barrier to political engagement," *Political Communication*, vol. 39, no. 6, pp. 697–714, 2022, doi: [10.1080/10584609.2022.2123073](https://doi.org/10.1080/10584609.2022.2123073).
- [11] K. Valaskivi, "Circulation of conspiracy theories in the attention factory," *Public Communication*, vol. 20, no. 3, pp. 162–177, 2022, doi: [10.1080/15405702.2022.2045996](https://doi.org/10.1080/15405702.2022.2045996).

- [12] M. Seeliger, S. Sevignani, "A new structural transformation of the public sphere? An introduction," *Theory, Culture & Society*, vol. 39, no. 4, pp. 3–16, 2022, doi: [10.1177/02632764221109439](https://doi.org/10.1177/02632764221109439).
- [13] A. Chadwick, J. Stanyer, "Deception as a bridging concept in the study of disinformation, misinformation, and misperceptions: Toward a holistic framework," *Communication Theory*, vol. 32, no. 1, pp. 1–24, 2022, doi: [10.1093/ct/qtab019](https://doi.org/10.1093/ct/qtab019).
- [14] A. Wanless, M. Berk, "The audience is the amplifier: Participatory propaganda," in *The Sage handbook of propaganda*, P. Baines, N. O'Shaughnessy, N. Snow (eds.), Los Angeles: SAGE, 2020, pp. 85–104.
- [15] N. Just, M. Latzer, "Governance by algorithms: Reality construction by algorithmic selection in the Internet," *Media, Culture & Society*, vol. 39, no. 2, pp. 238–258, 2017, doi: [10.1177/0163443716643157](https://doi.org/10.1177/0163443716643157).
- [16] H. Song, H. Gil de Zúñiga, H.G. Boomgarden, "Social media news use and political cynicism: Differential pathways through 'news finds me' perception," *Mass Communication and Society*, vol. 23, no. 1, pp. 47–70, 2017, doi: [10.1080/15205436.2019.1651867](https://doi.org/10.1080/15205436.2019.1651867).
- [17] J. Harambam, K. Grusauskaite, L. de Wildt, "Poly-truth, or the limits of pluralism: Popular debates on conspiracy theories in a post-truth era," *Public Understanding of Science*, vol. 31, no. 6, pp. 784–798, 2022, doi: [10.1177/09636625221092145](https://doi.org/10.1177/09636625221092145).
- [18] T. Harjunemi, "Post-truth, fake news and the liberal 'regime of truth': The double movement between Lippmann and Hayek," *European Journal of Communication*, vol. 37, no. 3, pp. 269–283, 2022, doi: [10.1177/02673231211046784](https://doi.org/10.1177/02673231211046784).
- [19] A. Jungherr, R. Schroeder, "Artificial intelligence and the public arena," *Communication Theory*, vol. 33 no. 1–2, pp. 164–173, 2023, doi: [10.1093/ct/qtad006](https://doi.org/10.1093/ct/qtad006).
- [20] U. Hasebrink, L. Merten, J. Behre, "Public connection repertoires and communicative figurations of publics: Conceptualizing individuals' contribution to public spheres," *Communication Theory*, vol. 33, no. 2–3, pp. 82–91, 2023, doi: [10.1093/ct/qtad005](https://doi.org/10.1093/ct/qtad005).
- [21] A. Bruns, "From 'the' public sphere to a network of publics: Towards an empirically founded model of contemporary public communication spaces," *Communication Theory*, vol. 33, no. 2–3, pp. 71–81, 2023, doi: [10.1093/ct/qtad007](https://doi.org/10.1093/ct/qtad007).
- [22] E. Esposito, *Artificial communication: How algorithms produce social intelligence*. Cambridge, MA: MIT Press, 2022.
- [23] A. Hepp, N. Couldry, "Necessary entanglements: Reflections on the role of a 'materialist phenomenology' in researching deep mediatization and datafication," *Sociologica*, vol. 17, no. 1, pp. 137–153, doi: [10.6092/issn.1971-8853/15793](https://doi.org/10.6092/issn.1971-8853/15793).
- [24] N. Schaetz, E. Gagrčin, R. Toth, M. Emmer "Algorithm dependency in platformized news use," *New Media & Society*, 2023, doi: [10.1177/14614448231193093](https://doi.org/10.1177/14614448231193093).
- [25] A. Hepp, *Deep mediatization*. London: Routledge, 2020.
- [26] D. Lupton, A. Watson, "Towards more-than-human digital data studies: Developing research-creation methods," *Qualitative Research*, vol. 21, no. 4, pp. 463–480, 2021, doi: [10.1177/1468794120939235](https://doi.org/10.1177/1468794120939235).

- [27] R. Braidotti, *Posthuman knowledge*. Cambridge: Polity, 2019.
- [28] J. Monforte, "What is new for new materialism for a newcomer," *Qualitative Research in Sport, Exercise and Health*, vol. 10, no. 3, pp. 378–390, 2018, doi: [10.1080/2159676X.2018.1428678](https://doi.org/10.1080/2159676X.2018.1428678).
- [29] D. Lupton, *Data selves*. Cambridge: Polity, 2020.
- [30] R. Braidotti, "A theoretical framework for the critical posthumanities," *Theory, Culture & Society*, vol. 36, no. 6, pp. 31–61, 2019, doi: [10.1177/0263276418771486](https://doi.org/10.1177/0263276418771486).
- [31] C. Pentzhold, A. Bischof, "Making affordances real: Socio-material prefiguration, performed agency, and coordinated activities in human-robot communication," *Social Media + Society*, 2019, doi: [10.1177/2056305119865472](https://doi.org/10.1177/2056305119865472).
- [32] T. Oliver, *The self-delusion: The surprising science of how we are connected and why that matters*. London: Weidenfeld and Nicolson, 2020.
- [33] N. Charter, *The mind is flat: The illusion of mental depth and the improvised mind*. New Haven: Yale University Press, 2019.
- [34] M. Coeckelbergh, *AI ethics*. Cambridge, MA: MIT Press, 2020.
- [35] J. Twomey, D. Ching, M. P. Aylett, M. Quayle, C. Linehan, G. Murphy, "Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine," *PloS ONE*, vol. 18 no. 10, pp. 1–22, 2023, doi: doi.org/10.1371/journal.pone.0291668.
- [36] V. Elliott. (July 27, 2023). *Big AI won't stop election deepfakes with watermarks*, *Wired*. [Online]. Available: <https://www.wired.com/story/ai-watermarking-misinformation/>. [Accessed: 10 November 2023].
- [37] T. Benson. (June 18, 2023). *Humans aren't mentally ready for an AI-saturated 'post-truth world'*, *Wired*. [Online]. Available: <https://www.wired.com/story/generative-ai-deepfakes-disinformation-psychology/>. [Accessed: 10 November 2023].
- [38] M. Cantor. (May 8, 2023). *Nearly 50 websites are 'AI-generated', a study says. Would I be able to tell?*, *The Guardian*. [Online]. Available: <https://www.theguardian.com/technology/2023/may/08/ai-generated-news-websites-study>. [Accessed: 10 November 2023].
- [39] W. Knight. (August 29, 2023). *"It costs just \$400 to build an AI disinformation machine," Wired*. [Online]. Available: <https://www.wired.com/story/400-dollars-to-build-an-ai-disinformation-machine/>. [Accessed: 10 November 2023].
- [40] N. Hagar, N. Diakopoulos, "Algorithmic indifference: The dearth of news recommendations on TikTok," *New Media & Society*, 2023, doi: [10.1177/14614448231192964](https://doi.org/10.1177/14614448231192964).
- [41] M. Burgess, M. (May 25, 2023). *"The security hole at the heart of ChatGPT and Bing," Wired*. [Online]. Available: <https://www.wired.co.uk/article/chatgpt-prompt-injection-attack-security>. [Accessed: 10 November 2023].
- [42] W. Knight. (August 1, 2023). *"A new attack impacts major AI Chatbots – and no one knows how to stop it," Wired*. [Online]. Available: <https://www.wired.com/story/ai-adversarial-attacks/>. [Accessed: 10 November 2023].

- [43] S. Alemohammad, J. Casco-Rodriguez, L. Luzi, A. I. Humayun, H. Babaei, D. Lejeune, A. Siahkoohi, R. G. Baraniuk, "Self-consuming generative models go MAD," *arXiv*, 2023, doi: [10.48550/arXiv.2307.01850](https://doi.org/10.48550/arXiv.2307.01850).
- [44] M. Bay, "Arendt in the metaverse: Four properties of extended reality that imperil *factual truth and democracy*," *Convergence*, vol. 29, no. 6, pp. 1698–1712, 2023, doi: [10.1177/13548565231199957](https://doi.org/10.1177/13548565231199957).
- [45] M. Hameleers, F. Marquart, "It's nothing but a deepfake! The effects of mis-information and deepfake labels delegitimizing an authentic political speech," *International Journal of Communication*, 17, pp. 6291–6311, 2023.
- [46] M. Holmstrom, "The narrative and social media," *Defence Strategic Communications*, vol. 1, no. 1, pp. 119–133, 2015, doi: [10.30966/2018.RIGA.1.7](https://doi.org/10.30966/2018.RIGA.1.7).
- [47] U.K.H. Ecker, "Why rebuttals may not work: The psychology of misinformation," *Media Asia*, vol. 44, no. 2, pp. 79–87, 2018, doi: [10.1080/01296612.2017.1384145](https://doi.org/10.1080/01296612.2017.1384145).
- [48] A. de Albuquerque, T. M. Oliveira, M. A. dos Santos Jr, R. Quinan, D. Mazur, "Coronavirus meets the clash of civilizations," *Convergence: The International Journal of Research into New Media Technologies*, vol. 28, no. 4, pp. 1198–1213, 2022, doi: [10.1177/13548565221105789](https://doi.org/10.1177/13548565221105789).
- [49] M. Andrejevic, *Automated media*. London: Routledge, 2020.
- [50] M. Butter, *The nature of conspiracy theories*. Cambridge: Polity, 2020.
- [51] G. Cosentino, *Social media and the post-truth world order*. London: Palgrave Macmillan, 2020.
- [52] A. Bruns, S. Harrington, E. Hurcombe, "'Corona? 5G? or both?': The dynamics of COVID-19/5G conspiracy theories on Facebook," *Media International Australia*, vol. 177, no. 1, pp. 12–29, 2020, doi: [10.1177/1329878X20946113](https://doi.org/10.1177/1329878X20946113).
- [53] N. O'Shaughnessy, "From disinformation to fake news: Forwards into the past," in *The Sage handbook of propaganda*, P. Baines, N. O'Shaughnessy, N. Snow Eds., Los Angeles: SAGE, 2020, pp. 55–70.
- [54] I. Kalpokas, "Influence operations: Challenging the social media–democracy nexus," *SAIS Europe Journal of Global Affairs*, vol. 19, no. 1, pp. 18–29, 2016.
- [55] I. Kalpokas, "Information warfare on social media: A brand management perspective," *Baltic Journal of Law and Politics*, vol. 10, no. 1, pp. 35–62, 2017, doi: [10.1515/bjlp-2017-0002](https://doi.org/10.1515/bjlp-2017-0002).
- [56] I. Kalpokas, "Social media: Mimesis and warfare," *Lithuanian Foreign Policy Review*, vol. 35, pp. 116–133, 2016, doi: [10.1515/lfpr-2016-0006](https://doi.org/10.1515/lfpr-2016-0006).
- [57] L. Schwaiger et al., "Mindsets of conspiracy: A typology of affinities towards conspiracy myths in digital environments," *Convergence: The International Journal of Research into New Media Technologies*, vol. 28, no. 4, pp. 1007–1029, 2022, doi: [10.1177/13548565221106427](https://doi.org/10.1177/13548565221106427).
- [58] I. Z. Baron, *How to save politics in a post-truth Era*. Manchester: Manchester University Press, 2018.
- [59] E. Bonetto, T. Arciszewski, "The creativity of conspiracy theories," *Journal of Creative Behavior*, vol. 55, no. 4, pp. 916–924, 2021, doi: [10.1002/jocb.497](https://doi.org/10.1002/jocb.497).

- [60] J.E. Uscinski, A.M. Enders, "What is a conspiracy theory and why does it matter?" *Critical Review: A Journal of Politics and Society*, vol. 35, no. 1–2, pp. 148–169, 2023, doi: [10.1080/08913811.2022.2115668](https://doi.org/10.1080/08913811.2022.2115668).
- [61] S. Altay, A. Acerbi, "People believe misinformation is a threat because they assume others are gullible," *New Media & Society*, 2023, doi: [10.1177/14614448231153379](https://doi.org/10.1177/14614448231153379).
- [62] I. Kalpokas, J. Kalpokiene, "Reimagining political competition for the age of data and algorithm," *Giornale di Filosofia*, vol. 2, pp. 33–46, 2022, doi: [10.7413/1827-5834013](https://doi.org/10.7413/1827-5834013).

Vulnerabilities of Web Applications: Good Practices and New Trends

Mateusz Nawrocki | Cracow University of Technology, Poland | ORCID: 0009-0007-5370-3497

Joanna Kołodziej | NASK – National Research Institute, Poland | ORCID: 0000-0002-5181-8713

Abstract

Web application security remains a critical challenge in mitigating vulnerabilities that expose sensitive data and systems to cyberattacks. This paper addresses the recent trends in the vulnerability of web applications to cyberattacks. It explores implementing and evaluating security mechanisms in web services guided by the Open Web Application Security Project's (OWASP) Top 10 framework. The OWASP analyser – a test application prepared to simulate the broken access control, Structured Query Language (SQL) Injection, and cross-site scripting (XSS) attacks – was executed in three realistic scenarios: web applications without any protection mechanism, essential safeguards, and advanced measures. The experimental results demonstrate the effectiveness of layered security strategies and highlight the best practices, such as role-based access control, secure cryptographic methods, and comprehensive logging. The analysis highlights the need to embed security throughout Web applications' implementation and use cycle. While advanced measures, such as encryption and real-time monitoring, increase resilience to sophisticated attacks, even basic practices can provide significant application protection if applied consistently.

Keywords

cybersecurity, vulnerability, web application, OWASP TOP 10

Received: 11.10.2024

Accepted: 20.12.2024

Published: 25.12.2024

Cite this article as:

M. Nawrocki, J. Kołodziej, "Vulnerabilities of web applications: Good practices and new trends," ACIG, vol. 3, no. 2, 2024, pp. 122–143. DOI: 10.60097/ACIG/199521

Corresponding author:

Mateusz Nawrocki, Cracow University of Technology, Poland; E-mail: mateusz.nawrocki@pk.edu.pl

 0009-0007-5370-3497

Copyright:

Some rights reserved

(CC-BY):

Mateusz Nawrocki

Joanna Kołodziej

Publisher NASK



1. Introduction

Web applications have become deeply embedded in various fields and aspects of functioning in the IT era. It isn't easy to consider modern e-commerce management systems, communications, entertainment, and banking and financial services without sophisticated, intelligent, and responsive web services, recently supported by artificial intelligence (AI). The Internet was recognized in the early 1990s as the sixth primary mass medium in civilisation's development. Thus, web applications have become the foundation for developing modern digital technologies. However, the ubiquity of digital applications makes them prime targets for cyberattacks. Security gaps and any vulnerability to external manipulation are exploited to steal data, users' identities, and, finally, to obtain specific financial benefits [1].

Security in IT refers primarily to ensuring the stability and resilience of various applications, systems, and data against unauthorised attacks that may result in illegal access to these resources. Over the past few years, this issue has been a frequent topic of commercial reports prepared for various institutions, from global agencies and government structures to scientific publications of interest mainly to the academic community. An example of such a publication is the work of Al-Ibrahim and Al-Deen [1], which describes the principal vulnerabilities of educational and research-related websites and services and methods to counteract the poisoning of content published there. The authors point out differences like these threats depending on the ownership structure of the university or school and the education profile. An example of a publication aimed at e-commerce environments is the report prepared by Thuraisingham et al. [2], which – in addition to threats and the most common attacks – describes tools to support methods of controlling access to the infrastructure and resources of a given company as well as secure systems for managing workflow in a company or an organisation.

Within the rapidly evolving IT sector, threats are also undergoing continuous transformation. As the amount of sensitive data available online increases, so does the need for tools to protect it. Effective data protection systems for online systems should work based on the following basic principles:

- Separation of databases from applications (installation on different servers)
- Encryption of data files and backups
- Widely used firewalls and other methods to limit access to sensitive data.

However, using even the strictest procedures and the most perfect tool does not provide a 100% guarantee of protecting data and resources from unauthorised access and use. For example, when identifying and analysing threats such as phishing, it is essential to remember that these attacks often take advantage of human naiveté and inattention by enabling unauthorised access to sensitive data [3]. Today, user behaviour and preferences are the weakest links in the security chain [4].

The Open Web Application Security Project (OWASP) initiative [5] has played a fundamental role in identifying and developing prevention methods for Web application security's highly complex thematic horizon. OWASP is not only a project but a global community that makes efforts to improve web application security. This community's main activities are identifying and compiling complex taxonomies, ranking threats, and developing strategies and guidelines for mitigating and eliminating security vulnerabilities in web applications. Every 3 years OWASP publishes Top 10 reports on the most critical vulnerabilities affecting the security of web applications, highlighting areas that require focused attention and the implementation of appropriate protective mechanisms [6].

The research presented in this paper aims to briefly analyse web application vulnerabilities and evolving trends in developing strategies for securing these applications by comparing and analysing the last two editions of the OWASP Top 10 reports from 2017 and 2021. Understanding the nature and sources of web application vulnerabilities to attack and manipulation is paramount in an era where web applications are integral parts of our digital lives.

Based on the latest OWASP 2021 report, an OWASP analyser (OA) tool was developed and installed to illustrate web application vulnerabilities to specific attacks. OA is a hybrid web application combining features frequently encountered in social media platforms, e-commerce websites, and content management systems (CMS). It comprehensively tests various vulnerabilities and security defences in a single web environment. Through deploying multiple layers of security, ranging from basic defence mechanisms (or lack thereof) to approaches from OWASP Top 10 recommendations, OA monitors exploitation of the specific vulnerabilities and assesses which defence mechanisms best mitigate or prevent possible attack scenarios.

The experiments conducted using AO were aimed at identifying critical vulnerabilities of web applications based on the guidelines

of the 2021 edition of the OWASP Top 10 report. Three application variants were implemented – from a version with no protection through an iteration containing basic security measures to a robust configuration using advanced defences. Penetration tests were conducted using popular security tools, such as Burp Suite [7], simulating Structured Query Language (SQL) Injection, and XSS and phishing attacks. The effectiveness of the defence methods used in the tests was evaluated, as was the difficulty level in bypassing each security measure. The experimental analysis concludes with recommendations and insights to raise awareness among developers and end users of the most prevalent cyber security threats.

The remainder of the paper is organised as follows. Section 2 outlines the main tenets of the OWASP Top 10 reports. The last two editions of these reports were compared, and a simple comparative analysis highlighted changes in the threat landscape and trends in the development of effective web application security tools. Section 3 describes the architectural model concepts, functions, and relationships between the main components of the OA application. The experimental analysis and results obtained are described at length in Section 4. Section 5 lists the most important web application security guidelines. The work concludes with Section 6.

2. OWASP Top 10 – A Review of 2017 and 2021 Editions

OWASP is a global non-profit organisation that brings together security experts and developers striving to improve the security of web applications. OWASP reports have become a roadmap for the focused community. They set trends in the security market for modern intelligent web services.

The first OWASP Top 10 report was released in 2003, with subsequent updates following approximately every 3 years: 2004, 2007, 2010, 2013, and 2017. The latest revision was published in 2021, suggesting that a new edition may soon be on the horizon. Each release reflects shifts in the threat landscape, incorporating new vulnerabilities and attack vectors that emerge alongside evolving technologies [8].

This section presents a brief comparative analysis of the last two editions of the OWASP Top 10 reports: the 2017 edition and the 2021 edition. The analysis focuses primarily on the evolution of web application vulnerabilities over just 3 years. Comparing these two OWASP reports shows how crucial the OWASP community is in shaping web application security practices.

2.1. Review of the 2021 Edition of OWASP Top 10

Below is a short review of the vulnerabilities in the 2021 edition of the OWASP report, which outlines the changes compared to the 2017 edition, along with the key threats and recommended defences.

1. **A01 – 2021: Broken access control.** It occurs when applications fail to enforce access restrictions properly and give unauthorised users access to the data. This vulnerability is the most serious web application security risk (5th position in the 2017 edition).
2. **A02 – 2021: Cryptographic failures.** It results from insufficient or improper use of cryptography, such as storing passwords in plain text or using outdated algorithms. Previously known as A03 – 2017: Sensitive Data Exposure, it often leads to sensitive data exposure.
3. **A03 – 2021: Injection.** It involves unvalidated user input reaching an interpreter (e.g. databases), leading to the execution of unintended commands. Defensive measures include parameterised queries and thorough validation of all inputs. The 2021 edition contains XSS.
4. **A04 – 2021: Insecure design.** A new category that highlights shortcomings in the early design stages, such as neglecting threat modelling or risk assessment.
5. **A05 – 2021: Security misconfiguration.** This covers a wide range of misconfigurations, like leaving default credentials unchanged, enabling debug modes in production, or inactive unnecessary features. It was A06 in the 2017 edition.
6. **A06 – 2021: Vulnerable and outdated components.** It addresses risks of running unsupported or outdated software components, libraries, and frameworks. Regular updates and dependency checks are critical for mitigation.
7. **A07 – 2021: Identification and authentication failures – formerly known as Broken Authentication.** This focuses on weak passwords, insufficient multi-factor authentication (MFA), or poor session management. Enforcing strong password policies and secure session handling is crucial.
8. **A08 – 2021: Software and data integrity failures.** This emphasises maintaining the integrity of application code and data, such as verifying software updates via digital signatures and using secure serialisation – a new category in 2021.
9. **A09 – 2021: Security logging and monitoring failures.** This category reflects the importance of logging and monitoring security events. Without proper logs or alerting mechanisms, attacks may go unnoticed for long periods.

10. A10 – 2021: Server-side request forgery (SSRF). It allows attackers to manipulate server-side requests, potentially accessing internal systems or sensitive data that would otherwise be restricted.

2.2. Comparison of the 2017 and 2021 Editions

The results of a short comparative analysis of both OWASP Top 10 editions are presented in Fig. 1.

The 2017 report outlined 10 critical threats: injection, broken authentication, and security misconfiguration. The main changes in the 2021 report can be defined as follows:

1. Three new categories

- a. *Insecure design* emphasises addressing security considerations during the initial design phase.
- b. *Software and data integrity failures* focuses on code and data integrity issues, including secure software updates and serialisation methods.
- c. *C. server-side request forgery (SSRF)* highlights vulnerabilities allowing attackers to manipulate server-side requests to access internal resources.

2. Renamed and merged categories

Particular vulnerabilities were combined or renamed to reflect better current cybersecurity challenges (e.g. *broken authentication* became *identification and authentication failures*).

2. Greater emphasis on secure design

The 2021 edition underlines the need to integrate security throughout the entire application lifecycle, rather than viewing it solely as an implementation concern.

A comparative analysis of the 2021 and 2017 versions of the OWASP Top 10 reveals several observations about changing trends in web application security vulnerabilities. These observations have substantial implications for security practitioners, developers, and organisations looking to strengthen their web application security measures. One finding is the continued presence of three threats that emerged in 2017. This means that injection, broken access control and cryptographic failure threats are still relevant and must be prioritised consistently [8]. The 2021 report updates the descriptions and scope of some vulnerabilities to account for the dynamics of changes in the development of modern web application architectural models.

Two new classes of vulnerabilities featured in the 2021 edition; “Software and Data Integrity Failures” and “Unsecured Design,”

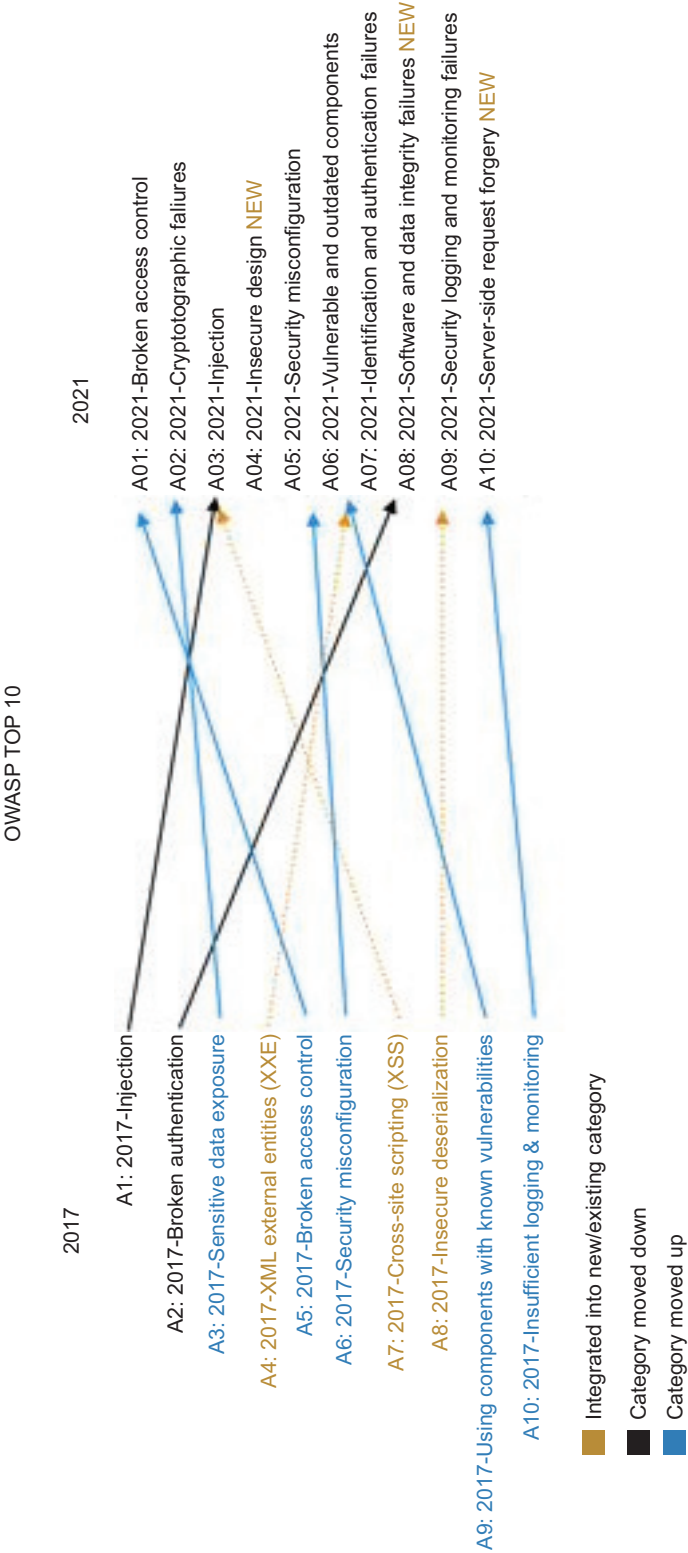


Figure 1. Comparison of vulnerability rankings in the 2017 and 2021 editions of OWASP Top 10 reports (based on the source [8]).

emphasise the importance of addressing security issues as early as a web application's design and implementation stage. The growing prevalence of data integrity in modern applications was also highlighted.

The simple comparative analysis conducted in this chapter has confirmed the dynamics of the development of Internet applications along with new threats. Many researchers refer to this phenomenon as a dynamic threat landscape. Security strategies need to evolve at a pace similar to the development of application development tools, and employers and organisations need to adopt flexible and adaptive strategies for web application security, staying abreast of the latest trends and vulnerabilities. Such adaptability is critical to effectively countering new and emerging threats.

3. OWASP Analyser

The original OA application was inspired by illustrating specific threat scenarios and the significant vulnerabilities of web applications to selected attacks, such as injection or XSS. The application is designed to demonstrate the consequences of these attacks in specific examples – this could be data theft or unauthorised access to an account. With the use of OA, it becomes possible to demonstrate robust defence mechanisms and raise awareness among users and developers about the critical importance of adequate security measures throughout the software lifecycle.

There are already some applications on the market with similar functionality. One example is Damn Vulnerable Web Application (DVWA) [9]. DVWA is a web application designed as a tool for learning and testing various security techniques, such as SQL Injection and XSS, in a controlled environment. It is particularly useful for beginners in penetration testing and vulnerability analysis. Another example is OWASP Juice Shop [10], which is used to simulate real-world security vulnerabilities. OWASP Juice Shop is a modern web application aimed at learning and testing skills in cybersecurity. The project supports development in application security by providing scenarios aligned with the OWASP Top 10. Another example is bWAPP [11], which is particularly useful for practicing and understanding over 100 web vulnerabilities, including those outlined in the OWASP Top 10, in a safe and controlled environment, making it an excellent tool for security enthusiasts, developers, and students to enhance their knowledge of web application security. Against

this background, OA stands out for its flexibility and easy adaptability to different threat landscapes, and meets the guidelines in the latest OWASP top 10 reports.

3.1. OWASP Analyser Architectural Model

Figure 2 presents a high-level diagram of the OA architectural model. This model consists of three main interconnected modules: a client-side interface module, a Python-based server, and an SQLite database.

3.1.1. Client Module

The client module was implemented using typical web technologies, including HTML, CSS, and TypeScript, combined with the angular and angular material frameworks. These tools were

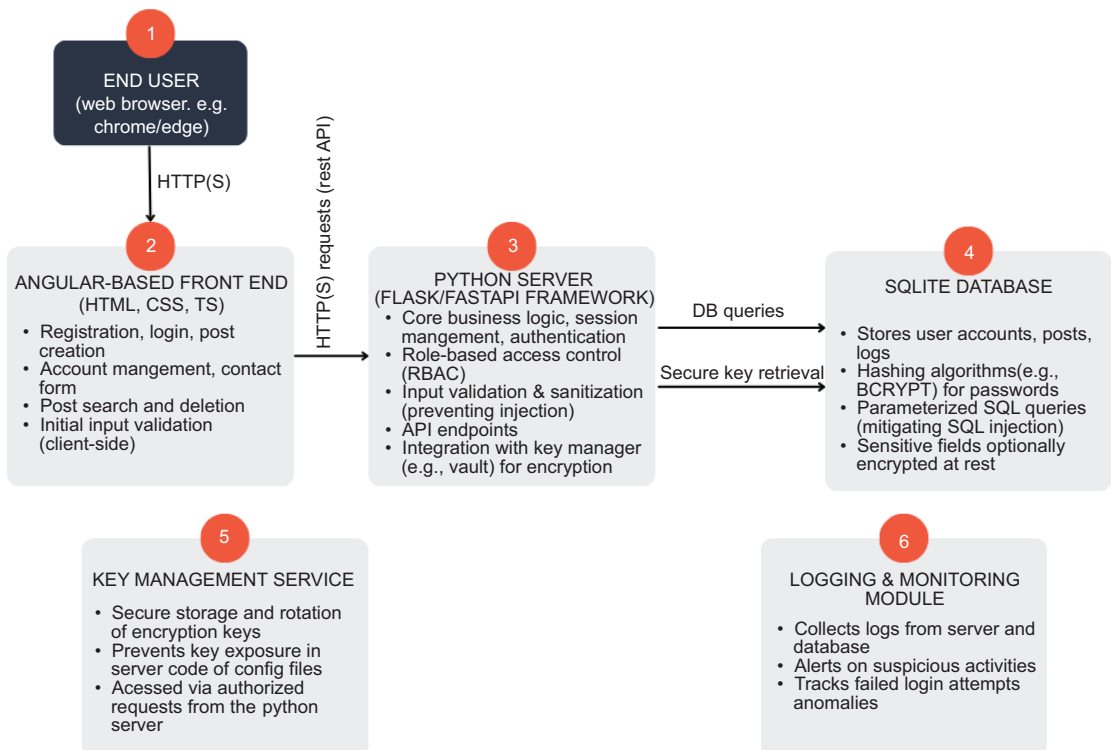


Figure 2. Architectural model of OWASP Analyzer. A high-level overview of six interconnected modules: client-side front end, Python server, SQLite database, key management, and logging & monitoring for secure operations.

used to develop a user-friendly responsive interface while maintaining flexibility for continuous improvements in security mechanisms. For example, TypeScript's strong typing system has facilitated accurate input validation and helped to minimise security vulnerabilities due to improper data handling. A concrete example is the login method, which uses strict type-checking to prevent malicious or invalid data from being processed, thus reducing the likelihood of injection attacks.

The client module contains the most commonly attacked functions, such as user registration, login, post creation, and account management. Initially, these functions were left unprotected to accurately simulate threats, such as XSS and cross-site request forgery (CSRF). This created a baseline environment, which was then used to test the effectiveness of the implemented security tools.

3.1.2. Server Module

Developed in Python, the server module handles basic tasks such as processing client requests, authenticating users, and interacting with the database. Security testing of the base OA versions revealed several security vulnerabilities, especially endpoints, without proper input validation and authorisation checks. For example, the /admin endpoint initially allowed access to unauthenticated users due to inadequate session token management.

In response, secure session management and robust backend access controls were implemented. Session tokens were dynamically generated using cryptographically secure random values and securely stored to prevent unauthorised reuse or tampering. Moreover, role-based access control restricted privileged operations, such as modifying user data, to authorised users only, reducing the likelihood of access controls being broken.

3.1.3. Database Layer

SQLite, a lightweight relational database, stores user credentials, posts, and other critical OA application data. Early iterations of the application faced significant risks, such as storing passwords in plain text. These issues were mitigated by adopting strong password hashing algorithms (such as bcrypt [12]) and ensuring that sensitive data remained encrypted during transmission and storage.

Additional safeguards included parameterised queries to protect against SQL Injection attacks. All user inputs were sanitised and processed with prepared instructions, preventing malicious actors from altering query logic or compromising database integrity.

3.1.4. Integrated Security Measures

Improvements in security mechanisms are methodically implemented and tested across all application layers. On the front end, input validation and client-side filters helped mitigate XSS attempts and invalid requests. On the server side, stricter authentication mechanisms, enforced API speed limits, and centralised logging were used to detect anomalies. The database encrypted critical fields, credentials were stored using secure hashing, and periodic audits were implemented to identify potential component vulnerabilities.

4. Experimental Analysis

The experimental analysis presented in this section shows the vulnerability of web applications according to the OWASP Top 10 2021 report. In the experiments, the designed and implemented OA application has the character of a social network. Social networks offer extensive configurations and options, which are easy to use for attackers. Nowadays, in the era of social media saturation, many people worldwide have several accounts on different sites. The application client is, therefore, simple and intuitive. OA in this implementation has the following functionalities:

1. login
2. Registering a new account with the application
3. Adding posts to the global board after completing the following information:
 - Title
 - Category
 - Location
 - Date of the event (if you create a post with an event)
 - Description
4. Deleting posts from the global board
5. Changing account settings – changing user name
6. Sending a contact form to the owner of web application
7. Searching for posts against criteria:
 - Title
 - Category
 - Location

- Start date
- End date

On the server module, addresses have been prepared waiting to send the appropriate request from the client module, which is checked and redirected to the database to perform the following operations:

1. Checking whether the logged-in user exists in the database, and whether his data is correct to allow logging into the application
2. Adding a new user to the database
3. Adding posts
4. Deleting posts
5. Editing posts
6. Log out
7. Checking authorisation
8. Deletion of authentication token data follows the logout process of the user; this provides additional security against unauthorised access
9. Checking whether a given user has authorisation to perform particular actions
10. Additional validation of data sent by the client application, should an attack be attempted
11. Encryption of data so that it does not leak during attacks
12. Providing information on the operation performed, whether it was successful, and what errors were intercepted
13. Sending a contact message to the application owner
14. Editing data regarding a particular account, error handling, for example, what if incorrect data is given,
15. Unblocking CORS for the particular address to which requests are made and those from which they are received,

The tests were conducted in the following three scenarios of OA configuration:

1. Basic configuration (without security measures)
In the initial phase, the application was tested without any security mechanisms. This identifies the main problems and vulnerabilities of the application related to lack of input validation, unsecured password storage, and unauthorised access to administrative resources. The test results in this scenario served as a baseline for subsequent testing phases.
2. Configuration with basic security measures
The second phase introduced minimal security measures, such as input validation, access restrictions to the administration

panel, and parameterised queries. These changes aimed to mitigate common threats, including SQL Injection and XSS attacks.

3. Configuration with advanced security

Comprehensive security mechanisms were implemented in the final phase of experimental analysis. These included role-based access control (RBAC), encryption of sensitive data, and automated session management. In addition, monitoring and logging mechanisms were integrated to reduce response time to potential attacks.

4.1. Results

The tests identified key weaknesses in application security and evaluated the effectiveness of various protection methods. The results of the experiments were interpreted in terms of the following five criteria in line with the OWASP Top 10 2021 report: flaws in access control mechanisms, vulnerability to SQL Injection attacks, vulnerability to XSS attacks, vulnerability to cryptographic errors and behaviour when logging, and monitoring tools are introduced.

4.1.1. Unauthorised Access Control

Initial tests showed that an unauthenticated user could access the administration panel through URL manipulation, for example, by appending/admin to the page address. This allowed unauthorised users to view sensitive data and potentially make system-level changes. For the basic configuration, all 30 attempts to access the admin panel without valid credentials succeeded, underscoring the severity of the vulnerability.

After implementing session management and enforcing server-side validation through the validation of authorisation tokens and user role verification, the vulnerability was successfully neutralised. In repeated tests in the new configuration, none of the 30 unauthorised access attempts failed, translating into a 100% success rate in blocking unauthorised logins.

Figure 3 shows the results of this experiment. It clearly shows the importance of combining session token validation with robust backend controls. Even if attackers try to manipulate URLs or inject forged session tokens, the server's RBAC mechanisms verify the authenticity of each request and privilege level. As a result, only legitimate and authenticated users with appropriate privileges are granted access to the/admin route, reducing the risk of data breaches or system misuse.

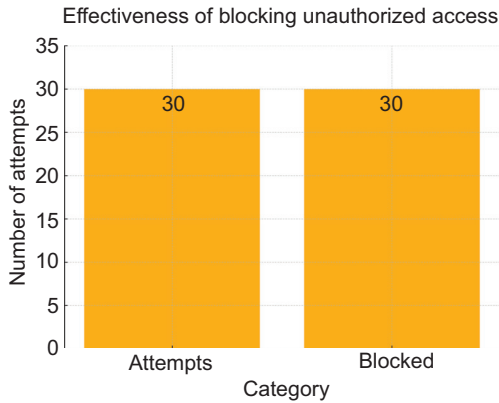


Figure 3. Effectiveness of blocking unauthorised access. Results demonstrate the significance of robust backend controls and session token validation in preventing unauthorised attempts.

4.1.2. SQL Injection

In the initial phase of the application's vulnerability to SQL Injection attacks, crafted queries (e.g. 'OR '1'='1') were injected, allowing attackers to retrieve sensitive data and unauthorised access to critical information without valid credentials. To address this problem, parameterised queries and additional input validation mechanisms were introduced, effectively sanitising user input before passing it to the database. As a result, all 25 SQL Injection attempts conducted in the follow-up tests were successfully blocked. In addition, logging functions were improved to track suspicious queries, thus enabling faster incident response and anomaly detection. The results of these experiments are presented in Fig. 4.

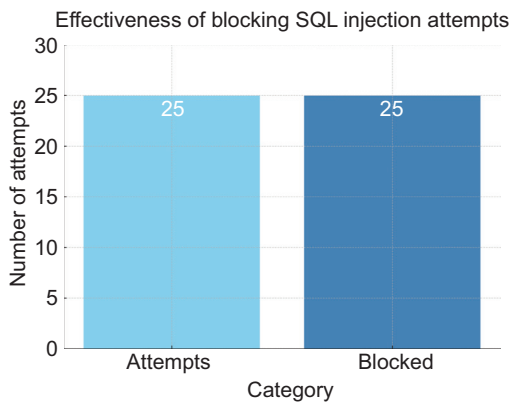


Figure 4. Effectiveness of blocking SQL Injection attempts. It highlights the reliability of security mechanisms in detecting and preventing all SQL Injection attempts.

4.1.3. Cross-Site Scripting

Cross-site scripting attacks primarily targeted a post-creation form, in which malicious scripts were injected to activate user accounts in browsers without their knowledge. In a basic configuration, these scripts are executed without any restrictions, posing a serious threat to data confidentiality and integrity.

Once input sanitisation and content escaping techniques were implemented, each of the 25 recorded XSS attack attempts was successfully neutralised (see Fig. 5). The security measures demonstrate that user-generated content is properly filtered and rendered as a plain text, rather than processed as executable code. Defence mechanisms are activated on the client side (to provide immediate feedback and prevent basic exploits) and on the server side (to validate and sanitise incoming data against more advanced payloads).



Figure 5. Effectiveness of blocking XSS attacks. It demonstrates the robustness of implemented security measures in successfully preventing all XSS attack attempts.

4.1.4. Cryptographic failures

Initially, the application stored passwords in plain text, which posed a serious security risk. The plain text credentials could be immediately exploited if an attacker gains access to the database. To address this vulnerability, the bcrypt hash algorithm was integrated, introducing a computational cost that makes brute-force attempts much more difficult.

In subsequent tests, brute-forcing passwords protected by bcrypt took more than 10 hours of continuous computation on standard hardware when configured with 10 rounds of hashing, as shown

in Fig. 6. This marked improvement illustrates the effectiveness of a robust hash function in protecting sensitive data. Salting and appropriate cryptographic parameters further reduced the likelihood of successful password cracking, ensuring that user credentials remain secure even during a partial database breach. The application strengthened its overall security status by adopting standard cryptographic practices and protected users from unauthorised account access.

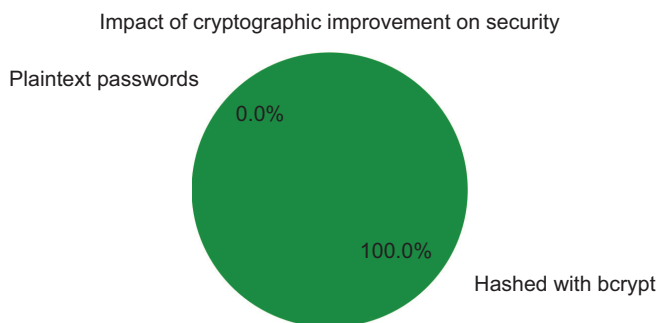


Figure 6. Impact of cryptographic improvements on security. This highlights complete transition from plain text passwords to secure hashing with bcrypt, ensuring enhanced data protection.

4.1.5. Login and Monitoring

The application failed to log login attempts or suspicious activity without activating logging and monitoring mechanisms, making detecting attacks or investigating incidents difficult. In the final configuration, event logging and real-time monitoring were implemented to capture critical security events, such as failed login attempts, SQL Injection attempts, and unusual user behaviour.

In most scenarios, these measures reduced response times from several hours (up to 240 minutes) to less than 15 minutes, enabling rapid intervention to stop threats. The results of the experiments are shown in Fig. 7, indicating the importance of continuous monitoring in implementing modern security strategies. By proactively analysing logs, setting up automatic alerts, and reviewing anomaly reports, organisations can respond quickly to potential breaches, thereby minimising damage and preserving the integrity of user data.

4.2. Summary of the Experiments

Security tests have demonstrated the key role of proactive measures in mitigating vulnerabilities commonly found in web

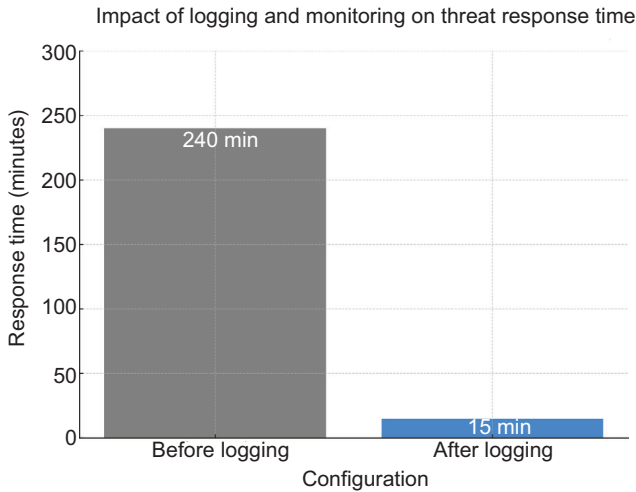


Figure 7. Impact of logging and monitoring on threat response time. It demonstrates a significant reduction in response time from 240 minutes to 15 minutes after implementing logging and monitoring mechanisms.

applications. Even basic security measures, such as input validation, session management, and parameterised queries, can significantly reduce the exposure of web services to threats, such as SQL Injection, XSS, or unauthorised access to data and the application itself.

The implemented advanced security mechanisms justified the concept of a modular architectural model for modern web applications. Role-based access control and robust session management have effectively neutralised access control security vulnerabilities in modular architectures. Implementing advanced cryptographic methods, such as the crypt hash method, has ensured sensitive information's integrity and data confidentiality. Logging and monitoring systems enabled rapid detection and response to suspicious activity, minimising potential damage from brute force and injection attacks.

The experimental analysis conducted led to the following guidelines for web application security:

- *Secure Application Life Cycle (SDLC)*: Security mechanisms must be integrated at every implementation and application life cycle stage, from design to deployment. Identifying potential threats through threat modelling and secure coding practices can prevent vulnerabilities from becoming web services.

- *Data validation and sanitisation*: These are fundamental to preventing injection attacks and ensuring data integrity. To ensure maximum effectiveness, input data validation should be enforced on both client and server sides.
- *Cryptographic standards*: Storing sensitive data, such as passwords, in plain text is a serious risk. Standard hash algorithms, such as bcrypt, and encrypting sensitive fields are essential to protect user information.
- *Access control policies*: Implementing detailed access control, including RBAC, allows users to access only the resources required for their roles. Backend validation should complement client-side controls to prevent circumvention.
- *Regular updates and dependency management*: Outdated software components can introduce vulnerabilities that can be exploited. Automated tools, such as dependency scanners, should be used to identify and regularly update unprotected libraries.
- *Comprehensive logging and monitoring*: Effective logging practices enable early detection of malicious activity. Monitoring tools should include real-time alerts for critical events, such as repeated logging failures.
- *Comprehensive logging and monitoring*: Effective logging practices enable early detection of malicious activity. Monitoring tools should include real-time alerts for critical events, such as repeated login failures or injection attempts, enabling rapid intervention.

5. Challenges and Future Trends

Despite marked improvements in vulnerability mitigation, especially in layered and modular architectures, organisations must constantly adapt to the rapidly evolving threat landscape. As technology advances, new types (vectors) of attacks and methodologies are emerging, requiring constant vigilance and innovation. Based on the theoretical and experiential analysis of the threats and vulnerabilities of modern web applications conducted in this work, the list of trends and key issues shaping the dynamics of broad changes in web application security, presented later in this section, has been defined.

5.1. Evolving Cyber Threats

Cybercriminals are rapidly refining their techniques, leveraging automation, social engineering, and AI-driven strategies to launch more sophisticated attacks. Traditional defences, such as basic firewalls or signature-based intrusion detection, may be insufficient against complex threats that adapt in real-time. This calls for

Table 1. Challenges and trends in web application security.

Challenge/trend	Description
Evolving Cyber Threats	Sophisticated attacks leveraging automation, AI, and real-time adaptation; requires anomaly detection and automated responses.
Integration with DevSecOps	Embedding security in CI/CD pipelines to identify vulnerabilities early and improve resilience.
Microservices and Containerisation	Granular controls and container isolation are needed to secure dependencies and configurations in distributed environments.
Zero-Trust Architecture	Continuous verification and dynamic access policies to minimise trust and lateral movement risks.
AI and Machine Learning (ML) for Defence	Using advanced machine learning for threat detection and response while addressing adversarial risks.
Regulatory Compliance and Data Privacy	Compliance with data privacy regulations to avoid fines and reputational damage.
Continuous Security Education	Training stakeholders to recognise threats and apply secure coding practices effectively.

advanced solutions that detect subtle anomalies, integrate threat intelligence, and automatically orchestrate responses to contain breaches before they escalate.

5.2. Integration with DevSecOps

The transition from traditional software development life-cycles (SDLC) to more agile and continuous delivery models has highlighted the need for DevSecOps [13], embedding security at every stage of development. Organisations can identify and remediate vulnerabilities earlier by automating security scans, code reviews, and penetration tests as part of the CI/CD pipeline. This approach reduces the likelihood of security issues persisting into production while ensuring faster release cycles and more resilient applications.

5.3. Microservices and Containerisation

Modern applications often adopt microservices architecture and containerisation (e.g. Docker, Kubernetes) for scalability and maintainability. However, each microservice and container introduces its own dependencies, configurations, and potential vulnerabilities. Securing these distributed environments requires granular access controls, robust container isolation, and regular updates of container images to prevent exploited or outdated components from compromising the entire system.

5.4. Zero-trust Architecture

A zero-trust model maintains that no user, device, or network segment is inherently trusted. Instead, continuous verification (e.g. using multi-factor authentication, dynamic access policies, and strict segmentation) becomes the standard. As remote work and cloud-based services expand, zero-trust frameworks help ensure that each request is rigorously validated, reducing the attack surface and limiting the lateral movement of adversaries once inside a network.

5.5. Artificial Intelligence and Machine Learning for Defence

While attackers leverage AI and machine learning to automate intrusion efforts, defenders can similarly employ these technologies for anomaly detection, threat intelligence, and real-time correlation of events. Advanced machine learning models can help differentiate legitimate user behaviour from malicious activities, significantly improving incident response. However, the risk of adversarial attacks (where attackers poison or manipulate machine learning models) remains an ongoing challenge that security teams must address.

5.6. Regulatory Compliance and Data Privacy

Growing awareness of data breaches and privacy violations has led to more stringent regulations, such as the General Data Protection Regulation (GDPR) [14] in the European Union and similar laws worldwide. Compliance requirements push organisations to adopt stricter security controls, encrypt sensitive data, and maintain detailed logs. Meeting these standards can be complex, but failure to do so exposes organisations to substantial fines and reputational damage.

5.7. Continuous Security Education

Human factors often represent the weakest link in the security chain. Social engineering, phishing, and credential theft rely on user error or lack of awareness. Regular training and awareness programs are vital for developers and end-users, helping them recognise threats, follow secure coding practices, and respond appropriately to security incidents. Ongoing education ensures that stakeholders can effectively navigate emerging threats and vulnerabilities.

6. Conclusions

This paper highlights the new trends and developments in the security of web applications. A list of the most important

vulnerabilities of these applications is published once every 3 years as the OWASP Top 10 report. This report includes a ranking of vulnerabilities indicating the most up-to-date at a given time and the most dangerous threats to users of web applications at all levels of their use. Section 2 presented a simple comparative analysis of the last two editions of the OWASP reports. This analysis showed how the threat landscape has changed in just 3 years and the tremendous need for flexible and responsive tools to prevent attacks and eliminate detected web application vulnerabilities.

Experimental results underscore the need to embed security throughout Web applications' implementation and use cycle. While advanced measures, such as encryption and real-time monitoring, increase resilience to sophisticated attacks, even basic practices can provide significant application protection if applied consistently.

Implementing automated security testing, coupled with ongoing education of developers and users on best practices, is essential to reduce the risk of losing data or sensitive information published online. In addition, regular audits and updates are the cornerstone of maintaining secure systems in an evolving threat landscape.

The future of web application security hinges on proactive integrated approaches that blend automation, zero-trust principles, and continuous monitoring. As organisations continue to embrace cloud computing, containerisation, and microservices architectures, DevSecOps practices have become indispensable, ensuring security measures are embedded at every development and deployment phase. By staying informed about the latest trends and adapting defences accordingly, stakeholders can better protect critical data and systems against the expanding spectrum of cyber threats.

References

- [1] M. Al-Ibrahim, Y.S. Al-Deen, "The reality of applying security in web applications in academia," *International Journal of Advanced Computer Sciences and Applications*, vol. 5, no. 10, pp. 7–16, 2014. doi: [10.14569/IJACSA.2014.051002](https://doi.org/10.14569/IJACSA.2014.051002).
- [2] B. Thuraisingham, C. Clifton, A. Gupta, E. Bertino, E. Ferrari (2002). Directions for web and e-commerce applications security [Online]. Available: <http://dx.doi.org/10.2139/ssrn.333682> [Accessed: Dec 20, 2024].
- [3] F. Salahdine, Z. El Mrabet, N. Kaabouch, "Phishing attacks detection a machine learning-based approach," *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York, NY, 2021, pp. 0250–0255.

- [4] Gartner, "What is Cybersecurity? Trends, Strategies, and Insights." [Online]. Available: <https://www.gartner.com/en/topics/cybersecurity> [Accessed: Dec. 20, 2024].
- [5] OWASP Foundation, "OWASP Foundation, the Open Source Foundation for Application Security." [Online]. Available: <https://owasp.org/> [Accessed: Dec. 20, 2024].
- [6] O.B. Fredj, O. Cheikhrouhou, M. Krichen, H. Hamam, A. Derhab, "An OWASP top ten driven survey on web application protection methods," in *Risks and security of internet and systems*, Lecture Notes in Computer Science, J. Garcia-Alfaro, J. Leneutre, N. Cuppens, R. Yaich, Eds. Cham: Springer, 2021, pp. 235–252.
- [7] PortSwigger Ltd., "Burp Suite," [Online]. Available: <https://portswigger.net/burp> [Accessed: Dec. 20, 2024].
- [8] OWASP Foundation, "OWASP Top Ten," [Online]. Available: <https://owasp.org/www-project-top-ten> [Accessed: Dec. 20, 2024].
- [9] Ryan Dewhurst, "Damn Vulnerable Web Application (DVWA)," [Online]. Available: <https://github.com/digininja/DVWA> [Accessed: Dec. 20, 2024].
- [10] OWASP Foundation, "OWASP Juice Shop," [Online]. Available: <https://owasp.org/www-project-juice-shop/> [Accessed: Dec. 20, 2024].
- [11] Malik Mesellem, "bWAPP: a buggy web application," [Online]. Available: <https://www.itsecgames.com/> [Accessed: Dec. 20, 2024].
- [12] Wikipedia, "bcrypt," [Online]. Available: <https://en.wikipedia.org/wiki/Bcrypt> [Accessed: Dec. 20, 2024].
- [13] IBM Corporation, "What is DevSecOps?" [Online]. Available: <https://www.ibm.com/think/topics/devsecops> [Accessed: Dec. 20, 2024].
- [14] European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council," [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng> [Accessed: Dec. 20, 2024].

Redefining Systemic Cybersecurity Risk in Interconnected Environments

Giacomo Assenza | Complex Systems & Security Lab, University Campus Bio-medico (UCBM) of Rome, Italy; the World Bank Group, US | ORCID: 0009-0007-4909-5775

Alessandro Ortalda | Brussels Privacy Hub, Vrije Universiteit Brussel (VUB), Faculty of Law and Criminology, Brussels, Belgium | ORCID: 0000-0001-9414-9938

Roberto Setola | Complex Systems & Security Lab, University Campus Bio-medico (UCBM) of Rome, Italy | ORCID: 0000-0002-8792-2520

Abstract

While the different entities that compose any socio-economic environment have always had a certain degree of interconnection, the evolving dynamics of cyberspace are intensifying their interdependence and shared reliance on the digital realm. This is giving rise to increasingly possible origins of systemic cybersecurity risk, potentially leading to scenarios where supply chains and essential services experience the rapid and widespread propagation of cascade events at unprecedented levels and velocities. If this interdependence is widely recognised and accepted (Section 2), the concept of systemic cybersecurity risk is still subjective and functional to the core mission of single components of a system (Sections 3 and 4), and this lack of common terminology prevents the community from adopting a shared posture to manage these risks. In this paper, we propose a workable and inclusive definition of systemic cybersecurity risk (Section 5). We then review relevant cybersecurity events arguing that while catastrophic episodes are still unseen, there are incidents that highlight systemic dynamics (Section 6). Finally, we review relevant diagnostic tools that have been developed to address systemic cybersecurity risks and

Received: 01.05.2024

Accepted: 07.08.2024

Published: 29.08.2024

Cite this article as:

G. Assenza, A. Ortalda, R. Setola "Redefining systemic cybersecurity risk in interconnected environments," ACIG, vol. 3, no. 2, 2024, pp. 144–169. DOI: 10.60097/ACIG/192119

Corresponding author:

Giacomo Assenza,
Complex Systems &
Security Lab, University
Campus Bio-medico
(UCBM) Rome, Italy; the
World Bank Group, US;
E-mail: giacomooassenza@
gmail.com

 0009-0007-4909-5775

Copyright:

Some rights reserved:
Publisher NASK



discuss their limitation as well as opportunities for future research (Section 7). We conclude by highlighting that systemic cybersecurity risk is, by definition, a shared risk, thus developing a common understanding is the starting point to endorse coordinated mitigations at system level.

Keywords

risk assessment, risk management, cybersecurity, systemic cybersecurity risk

1. Introduction

Ongoing evolutions of cyberspace dynamics have amplified the attention around systemic cybersecurity risks. The rapid integration of Information and Communication Technology (ICT) into societal functions has come together with a market concentration of products and services which has made the different entities constituting the socio-economic environment increasingly interconnected and dependent on shared infrastructure and common providers. In such a context, there is a growing concern that even single failures can spread across a system, leading to scenarios where supply chains and essential services could experience rapid and widespread cascading events at unprecedented scales. While most cybersecurity events traditionally have a narrowly defined set of victims [1], recent studies provide empirical evidence of an increased prevalence, scale, and impact of cyber-related incidents [2, 3]. Furthermore, recent episodes have demonstrated how failures can affect multiple entities simultaneously. For example, in May 2023 the exploitation of a vulnerability in the firewall system recommended by the industry body and adopted by most energy operators in Denmark led to 22 companies being compromised, with several of them forced to go into island mode operation [4]. In the report analysing the incident, it is highlighted that Denmark has a highly decentralised energy system composed of many small companies, which makes the sector fairly resilient in case of a single disruption. However, a situation of ‘systemic vulnerability – where the same vulnerability is exploited across companies’ can create a potentially critical situation [5]. This event is just one of the last of a series of episodes, such as WannaCry, NotPetya, SolarWind, and Log4j (and more recent ones like the CrowdStrike incident), which have demonstrated how failures can propagate across complex supply chains, emphasising how their reliance on shared infrastructure products and services, concentrates risk into an unknown number of critical nodes.

Despite the growing concern surrounding systemic cybersecurity risk, the underlying problems and potential solutions seem to remain unseizable and poorly understood. The concept of systemic cybersecurity risks results subjective and ambiguous in both literature and the community of cybersecurity practitioners, and so are the existing tools and methodologies for identifying and measuring sources of this type of risk. Currently, there is no shared terminology, and there is little agreement not only on what constitutes a systemic cybersecurity risk but also on the granularity at which a system can be defined (operator, sector, countries, or supra-national level), with existing definitions being functional to the mission of the entity defining them. This has so far hindered the development of a unified approach to managing these risks. The identification of what can be defined as a systemic cybersecurity risk is not just an academic exercise but it is seminal to understand how systemic dynamics affect cyberspace and consequently devise appropriate risk mitigation policies and incident response procedures. Systemic cybersecurity risk is, by definition, a shared risk, thus developing a common understanding is the starting point to endorse coordinated actions at the system level, both in terms of policies and operational capacities.

In this article, we first explore the broader concept of systemic risk and its roots in the financial sector (Section 2). Then, we turn to the existing approaches to defining and dealing with systemic cybersecurity risk, highlighting how these result in ad hoc and uncoordinated strategies. In particular, we briefly outline the existing interpretations, and argue that (i) currently systemic cybersecurity risk is a 'contextual' concept, with its definition heavily influenced by the specific mandate of the involved entities; (ii) existing approaches consider the systematicity of cybersecurity risk primarily on the impact that they may have, with limited attention given to the underlying dynamics that give rise to such risks (Sections 3 and 4). We then propose a comprehensive and flexible definition of systemic cybersecurity risk that can be applied at different levels of granularity, providing a common foundation for understanding and addressing the issue (Section 5). Subsequently, we apply our definition to review relevant case studies, arguing that while catastrophic cybersecurity incidents are still unseen, several cybersecurity events highlight systemic dynamics (Section 6). Finally, we review some of the diagnostic tools and methodological frameworks that have been developed, discussing how these efforts are undermined by a general lack of data, a partial and uneven application of methodologies, and a general resistance from operators to share information (Section 7).

2. The Emergence and Evolution of ‘Systemic Risk’ as a Concept

The concept of systemic risk emerged in the field of finance and economics, with some of the earliest references dating back to the aftermath of the Great Depression in the 1930s [6] when economists and policymakers began to recognise how the failure of individual entities, such as banks and financial institutions, could affect the entire economic and financial system. However, within the literature, more structured definitions of systemic risks started to appear only in the 1990s. The concept gained even further prominence during the 2007–2008 global financial crisis, when the collapse of major financial institutions and interconnectedness of financial markets highlighted the potential for shocks to cause far-reaching financial and economic downturns [7, p. 315, 8].

Definitions adopted between 1988 and 2014 by academics and banking institutions [9]¹ identify the following features of systemic risks: (i) *scale of the phenomenon*: systemic risks affect a large part of a system; (ii) *contagion effect*: due to the interdependencies and interconnectedness among its components, systemic risks have the potential to trigger a cascading series of adverse events spreading across the entire system; and (iii) *system failure*: systemic risks have the potential to impair the functioning of the system itself.

Overtime, the understanding of systemic risk has evolved, and its application expanded from the economic and financial perspectives of the early days to other disciplines and areas of study. Scholars and practitioners begun to approach the issue with the goal of understanding the dynamics of complex and cross-sector supply chains and the potential for widespread disruptions as a consequence of the interconnectedness and interdependence of infrastructures, processes, and services across the globe [10–14]. In particular, the exposure of society to systemic risks has been amplified by what is known as the information revolution [15]. Already in 1997, the US Presidential Commission on Critical Infrastructure Protection (PCCIP) concluded that the country was so reliant on ICT infrastructure that the government had to frame it within the broader ‘national security focus’ to address the impacts that would result for the entire nation in case of its disruptions [16]. Since then, hyperconnectivity, digitalisation, widespread deployment of Internet of Things, adoption of readily available cloud technologies, and, more broadly, the pace and reach of technological innovation have contributed to shaping a quickly evolving and interdependent environment. This environment makes it more difficult for

1——In Smaga (2014) [9], systemic risks are defined as ‘the risk that a shock will result in such a significant materialisation of (e.g. macro-financial) imbalances that it will spread on a scale that impairs the functioning of the financial system and to the extent that it adversely affects the real economy (e.g. economic growth)’.

operators to pursue business continuity, because they often have to rely on goods or services provided by other parties [17].

While this rapid innovation is bringing benefits in terms of efficiency and reach of operations, it is also introducing structural aspects that magnify the potential for risks. First, the interdependencies in the ICT ecosystem are growing in number and complexity, with related risks going beyond the mere technical aspect [18, 19]. As a result, both policy-makers and operators struggle with understanding the ‘intricate and interlocking dependencies’ [20], both upstream and downstream [21]. This often translates into inadequate risk management practices [22]. Second, the market concentration of digital services, where stakeholders often rely on similar – when not the same – technologies, infrastructures, services, and providers, implies that when these fail, the impact may affect a large number of assets and organisations [23]. Third, the growth of hacking capabilities and their availability through models, such as the Hacking-as-a-Service one, makes it easier and cheaper for malicious actors to operate [24]. Especially the large number of potential targets that can be hit with a single capability – see the above-mentioned feature (ii) of systemic risks – makes it appealing to attackers from a cost-benefits perspective.

These rapid developments have been largely acknowledged by the security community, which has increasingly focused on the structural vulnerabilities of societal functions [25–28] and has started to formulate the concept of ‘systemic cybersecurity risk’. However, there is little agreement on what these risks are, how to manage them, and even if they have ever materialised.

—— 3. Systemic Cybersecurity Risk is a Contextual Concept

When it comes to systemic cybersecurity risks, most of the academics and practitioners have kept the economic and financial perspective [29–32]. They refer to these risks as a subset of systemic financial risks where a cybersecurity event on systemic entities may lead to spillover effects. For example, the European Systemic Risk Board (ESRB) defines systemic cyber-related incidents as those occurring ‘in the financial sector’ and which could cause ‘serious negative consequences for the internal market and the real economy’ [33]. Similarly, the European Central Bank provides an understanding of systemic cyber risks within the broader context of macro-financial perspectives. Accordingly, systemic risks should be assessed by looking at the following two dimensions: (i) the cross-sectional dimension, which relates to how the risk propagates within the financial system;

and (ii) the time-related dimension, which looks at the dynamic evolution of financial stability risks over time and consider the procyclical build-ups of financial fragility [34]. On the other hand, other authors and practitioners have developed more comprehensive conceptualisations of systemic cybersecurity risks, which include aspects such as safety and security. These broader approaches are not limited to the financial sector, but apply to all sectors [1, 35]. For example, according to the World Economic Forum (WEF), systemic cybersecurity risk is ‘the risk that a cyber event at an individual component of a critical infrastructure ecosystem will cause significant delay, denial, breakdown, disruption, or loss, such that services are impacted not only in the originating component but consequences also cascade into related (logically and/or geographically) ecosystem components, resulting in significant adverse effects to public health or safety, economic security, or national security’ [36]. This approach entails a more inclusive concept that considers the goals of societal (and not only economic) wellbeing, and which therefore is extended to all the critical functions of society.

Another point of view that can be adopted to look at systemic cybersecurity risk concerns the level of granularity at which a ‘system’ is perceived. A system can be defined as a collection of interrelated and interconnected elements or components that work together to achieve a common purpose or goal² [37, 38]. Building on this definition, transnational processes, countries, sectors, societal functions, single operators, or even circumscribed sections of a corporate ICT landscape can all be characterised as systems [1]. Therefore, under this assumption, the adjective ‘systemic’ can assume different meanings depending on the point of view. For instance, while the WEF definition takes an international perspective, the 2021 Systemic Cyber Risk Reduction Venture – established by the US Cybersecurity and Infrastructure Security Agency (CISA) – adopts a national perspective, focusing on understanding how ‘cyber risks or incidents in individual pieces or components of National Critical Functions (NCF) could create far-reaching cascading impacts, leading to system-wide functional degradation or failure’ [39]. CISA’s understanding of ‘system’ corresponds to the United States as a country, and therefore its point of view on systematicity is nationally centred. In fact, it includes the risks that might affect the recognised NCFs³ (e.g., the provision of medical care, distribution of electricity, etc.), but it disregards the impacts that can manifest at the international level or be suffered by other countries. The 2023 US National Cybersecurity Strategy further emphasises this nation-centric view, highlighting the importance of addressing systemic risks to make the US digital ecosystem – clearly spelled out as ‘our digital

2——These elements can be tangible entities, such as physical objects or processes, as well as intangible entities, such as concepts or information flows. The interactions and relationships between the components of a system lead to emergent properties or behaviours that may not be evident when considering each component in isolation.

3——A list of 55 NCFs is available here: <https://www.cisa.gov/topics/risk-management/national-critical-functions>

ecosystem’ – resilient [40]. Going further down the abstraction scale, there is a well-established perspective that frames systemic cybersecurity risks within the context of enterprise risk management. In 2019, the Digital Director Network (DDN) released the DiRECTOR™ risk framework to help corporate boards and management teams to manage systemic risks in ‘complex digital business systems’ [41]. This framework defines systemic risk as the risk that a component’s failure in a corporate digital system propagates and escalates, putting the entire organisation at stake [42].

These definitions present significant differences but share the idea that systemic cybersecurity risks materialise after cybersecurity events that produce digital and physical damages, and create cascading effects across the system, with potentially significant disruptions. This perspective is rooted in the interdependence of functions and the importance that ICT has in modern systems. It looks at how widespread the impact of the cybersecurity risks is and considers this as the determinant variable to categorise a cybersecurity risk as ‘systemic’. However, this approach does little to determine the dynamics producing them. In other words, any cybersecurity risk with ‘far reaching cascading impacts’ [39] or ‘cascade into related (logically and/or geographically) ecosystem components’ [36] would be considered systemic. This blurs the different categorisations, on the one hand, between systemic cybersecurity risks and high-impact cybersecurity risks, and, on the other hand, between systemic cybersecurity risks and systemic risks more broadly. In fact, the widespread cascades of a cybersecurity event might be caused by physical or logical interdependencies, rather than digital or cyber ones, which entail that effective risk management measures are not necessarily driven by cybersecurity considerations.

A different and relatively new approach to typify systemic cybersecurity risk comes from the insurance industry. In recent years, insurance companies have been increasingly vocal about systemic cybersecurity risks, claiming that these challenge the sector’s capacity to provide adequate insurance coverage [43]. From their perspective, risks are systemic when they become uninsurable due to the massive losses that would arise from the interconnections among clients, sectors, and locations, as well as the difficulties of modelling and hedging [1]. For instance, the insurance company AIG defines as ‘systemic’ those risks that are ‘capable of impacting many companies at the same time’ [44]. Under this interpretation, insurers adopt a different definition of ‘system’: not anymore a group of elements working together towards a goal, but simply the group of entities that would be eligible to receive compensation in

case of cyber events. The apprehension pertains to the conceivable scalability, wherein a solitary incident could concurrently impact numerous companies, resulting in substantial interconnected liabilities for insurers. For instance, in the case of damages affecting cloud computing platforms employed by a large number of clients, the insurer would be compelled to settle claims for all its policyholders concurrently with evident economic losses [45]. For insurers, a particular type of systemic cyber risk relates to the so-called ‘cyber-war’ or, more generally, state-sponsored hacks, which, due to their high potential costs, most insurers are deciding not to cover [46]. For instance, Lloyd’s of London requires insurance policies to have an explicit exemption for state-backed computer network operations [47]. This approach could undermine trust and reliance on insurance instruments, as it creates uncertainty about the possibility of getting coverage where it is needed the most. In fact, not only is attributing a cyberattack, let alone identifying one as an act of war, a complex, multilayered, and ultimately political exercise [48, 49], but it is also well beyond the scope of insurers. Despite these complexities, insurers are trying to pursue this interpretation in practice, as shown by NotPetya and the consequent dispute between the US food company Mondelez and the Swiss insurance company Zurich (further analysed in Section 6).

Finally, a minoritarian interpretation of systemic cybersecurity risk examines it from the perspective of technological standardisation and adoption. In a sense, this is similar to the issue arising from the interconnectedness that characterises today’s systems that have been referred above. However, this conceptualisation does not focus on the cascading effect that an event might have. Rather, it looks at the fact that incidents involving certain technologies that are widely shared have near-instantaneous effects on a large surface, making traditional redundancy measures ineffective [50]. In a conventional non-cyber system, redundancy serves as a risk-reduction strategy. This is built on the assumption that not all systems fail simultaneously. However, in the realm of cybersecurity, this assumption does not necessarily hold true, as vulnerabilities, if exploited, might simultaneously affect all replicas. The SolarWinds episode (further analysed in Section 5), as well as the event in the Danish power sector, serve as prominent examples of this dynamic [50, 51].

4. Have Systemic Cybersecurity Events Occurred?

In Section 3, we presented different definitions of systemic cybersecurity risk, highlighting how these are highly context-related

and how they can be driven by subjective considerations. These aspects add complexity to the ongoing efforts to establish shared terminology for this evolving concept. Similarly, the lack of a common understanding prevents the community from organically identifying when and if systemic cyber risk has ever materialised. Many agree that while systemic cybersecurity risks are concrete, one of the main challenges related to understanding and managing them is the lack of data and case studies. For example, in 2019, the EastWest Institute asserted that no cybersecurity incidents had ever qualified as systemic [52]. To date, catastrophic cybersecurity events, which are likely to be unanimously labelled as systemic are still unseen [1], but the existing unclarity in the terminology and definitions creates substantial challenges in identifying if and how potential systemic dynamics have accompanied less evident, but still significant events.

For example, the 2021 Colonial Pipeline hack had a significant impact, but concentrated in the US energy sector. The incident forced the Colonial Pipeline, a crucial fuel transport system, to suspend operations for a week. This disruption led to widespread fuel shortages and price spikes along the East Coast, affecting numerous states and prompting panic buying. The Colonial Pipeline moves approximately 45% of the fuel supply for the East Coast, which made its shutdown particularly impactful. The incident resulted in an estimated 5500 gas stations running out of fuel, and the national average gas price saw an increase of around 8 cents per gallon in just 1 week [53]. According to some of the definitions analysed above, this episode could be seen as presenting systemic characteristics. It did have an impact in terms of price reaction and destabilised volatility [54], and it did disrupt one of the so-called NFCs categorised as systemic [1]. However, services were restored relatively quickly, the long-term impacts of this episode were limited, as well as its cascades on other sectors and countries, which would in turn undermine the categorisation of this incident as systemic under certain definitions of it, like the one from the WEF.

Similarly, classifying an event as systemic depends on the level of granularity at which a system is defined. WannaCry, for instance, was a 2017 ransomware that affected over 200,000 computers across 150 countries, with a specific concentration in the UK National Health Service (NHS). Within the NHS, it severely impacted 81 out of 236 NHS trusts, resulting in the cancellation of approximately 19,000 medical appointments. Also, the financial toll was significant, with the NHS estimated to have spent around £92 million in direct costs and lost revenue due to the hack [55]. Yet the impact

was significantly concentrated within the UK healthcare supply, with limited consequences on the delivery of the service globally. If analysed through a national/sectoral security-based framework, WannaCry is likely to be categorised as a systemic event, but the same label would be more difficult to apply from global or regional perspectives. Also, despite the significant loss of revenues and recovery costs, the event was far from resulting in economic or financial instability, which entails that financially focused definitions would disregard this incident as systemic.

On the other hand, the impacts from other episodes were severe enough to be considered systemic but distributed enough to elude this categorisation from national security-based framework. The 2017, NotPetya ransomware, which the White House stated to be the ‘most destructive and costly cyber-attack in history’ [56], had a substantial impact on various organisations across countries. The incident’s total cost to businesses worldwide has been estimated to be in the range of \$10 billion [57] and is reported to have affected countless machines around the world, from hospitals in Pennsylvania to a chocolate factory in Tasmania, affecting multinational companies, including FedEx’s European subsidiary TNT Express, the pharmaceutical giant Merck, French construction company Saint-Gobain, food producer Mondelez, and manufacturer Reckitt Benckiser, inflicting nine-figure costs in each case. One of the hardest-hit industries was shipping, with Maersk, a global shipping company, forced to suspend operations in 17 terminals around the globe [58], reporting losses of around \$300 million. The NotPetya incident also affected the insurance market. For instance, the refusal of Zurich Insurance Group to pay a \$100 million claim from food company Mondelez arguing that the stipulated policy was not liable to cover ‘warlike actions’ [59], led to a dispute between the Swiss and American companies. Eventually, the insurance company covered most of the damages created by NotPetya, but this created a precedent which resulted in industry-wide effort to update insurance policies with war exclusion clauses [47]. While from an international, sectoral, and corporate standpoints, the systemic element of NotPetya is difficult to deny, its impacts were distributed across many stakeholders globally to the extent that, except for Ukraine, where the effects were particularly manifest [57], in no countries the damages were so significant to be considered a national security issue, thus a systemic event from a national standpoint.

Existing impact-oriented definitions of systemic cybersecurity risks are functional, which means that the identification of this type of risk is largely subjective and dependent on the mission or

perception of entities at play. This creates challenges in studying, understanding, and addressing this phenomenon.

5. An Inclusive Definition

As described in the previous section, systemic cybersecurity risk is a highly contextual concept. This makes it difficult for the community of stakeholders to collaborate and effectively manage it. To address this challenge, a shared terminology or, at the very least, a mutual understanding must be developed. In this section we propose a definition of systemic cybersecurity risk, which tries to create common ground among different stakeholders. Following is the proposed definition:

A risk is to be considered as systemic cybersecurity risk when, within the context of the system under analysis, has the potential to initiate a cybersecurity event (trigger) that can spread over a number of other ICT parts or functionalities of the system (circuit) that is sufficient to create changes to the system (impact).

This definition takes an inclusive approach. The *trigger* refers to all the events that might lead to losses of confidentiality, integrity, or availability of information, data, or information (or control) systems [60, 61]. This includes events occurring through digital (such as malwares, ransomwares, distributed denial-of-services [DDoS], software failures, etc.) and physical (such as destruction or impairment of hardware, natural disaster) mean.

The *circuit* refers to the systems or networks of ICT assets, components, or infrastructures through which an initial trigger propagates. This is irrespective of the extension or surface of the system, meaning that the concept of circuit can be applied at different scales, as a system can comprise from a single entity to multiple entities distributed across sectors, countries, and regions. In order to differentiate systemic cybersecurity risks from broader systemic risks, the circuit relates to cyberspace only. This excludes physical or logic cascades that extend beyond the perimeter of ICT systems (e.g., a slowdown in the supply of healthcare services due to a shortage of goods due to a cybersecurity event in the provider of these goods). In fact, these types of cascades, while extremely relevant in a context of great interdependency between assets, do not necessarily require cybersecurity mitigations, which position them beyond the scope of this definition and related policy measures. It is also important to differentiate the circuit from the supply-chain

and related risks, the latter being a narrower concept referring to the people, processes, and technologies associated with the delivery of services from one entity to another [62].

Finally, *the impact* is intended as all the disruptions that may introduce changes to the system. This language suggests first that the impact implies a broader perspective than the mere economic-financial one suggested by some definitions adopted by experts (see Section 2). In fact, if it is likely that a catastrophic cyber-related event can affect the financial environment, this is not an essential condition for systematicity because, as we have established, systematicity is not a measure of the extensions but a measure of the perimeter within which a risk exists and materialises its impact. Second, while this definition recognises that the effect of a systemic cybersecurity risk is larger with respect to the generating trigger, it does not tie the idea of systematicity to high-impact events. Even if rare, there might be scenarios of cybersecurity events presenting systemic dynamics, which nonetheless did not affect aspects, such as national economy or security, and did not result in catastrophic or severe incidents. For instance, the Stuxnet malware self-replicated, infecting thousands of machines worldwide regardless of their operating system version, but it was designed to release its payload only in the nuclear power plant in Natanz [63, 64]. This means that, while the circuit in which the malware spread was extensive, the actual impact was circumscribed to a single operator with effects that resulted to be far from catastrophic.

6. Dynamics of Systemic Cyber Risks

The lack of an agreed upon definition translates into a lack of common taxonomies to categorise systemic cybersecurity risks and related events. This is exacerbated by a paucity of case studies. In this section, we adopt our definition outlined above to review relevant events and highlight how, even though none of them resulted in catastrophic effects, they still show systemic dynamics that can be helpful to understand, thus address, systemic cybersecurity risks. Systemic cybersecurity risk manifests in various forms and can be classified in multiple ways. In the following paragraphs, we analyse three different dynamics in which a trigger spreads across a circuit causing impacts. In particular, we identify *top-down*, *distributed*, and *independent* dynamics [1].

In a *top-down dynamics*, even a single event disrupting a critical component within a system has the potential to trigger a chain reaction that progressively influences a widening array of interdependent

entities. For example, if a critical asset of the Internet infrastructure fails, such as a submarine fibre optic communication cable (SCC), an Internet Exchange Point (IXP), or Domain Name Service (DNS), businesses and services operating over the Internet would be affected by the disruption and might be unable to deliver their services in a far-reaching domino effect. For example, SCCs handle 98% of the global traffic, and despite redundancies being available for most countries, there are episodes showing significant impacts of potential disruptions [65]. In 2015, in the archipelago of the Northern Marianas, the only available submarine cable was severed, cutting off the island from broadband traffic for days [66]. Impacts included a loss of access to the Internet and the collapse of communications, with disruptions in critical services (health, tourism, education, etc.), with estimated damages amounting to US\$21 million [67]. Similarly, data shows significant impacts triggered even where countries have redundancy systems and multiple alternative cables [68]. Other examples show the top-down dynamic that damages in the Internet infrastructure might cause. In 2016 Dyn, a major DNS provider in the United States, fell victim of a massive DDoS campaign launched by the Mirari Botnet that overwhelmed its servers with an unprecedented amount of traffic [69]. While the incident did not take down the Internet, caused catastrophic impacts, or affected the real economy, it did result in substantial disruptions and emphasised the ‘systemic role’ that single pieces of the Internet infrastructure play in maintaining the stability and availability of online services. The Dyn disruption resulted in a ‘massive East Coast Internet outage’ [70], and service disruptions for many major websites and online services that relied on Dyn’s DNS services. Popular services (like Twitter, Netflix, Reddit, Spotify, Airbnb, GitHub, Paypal, and more) were affected, either experiencing slow load times or becoming completely inaccessible for users not only in the United States but also in Europe and different parts of the world. Systemic cyber risks might also materialise following physical triggers; for instance, in 2019, a malicious fire in an Italian rail transformer room caused the unavailability of train data and information, which eventually caused significant delays and service suspensions [71]; or as part of broader systemic events, such as when the extreme weather caused a power outage in Gambia, which, in turn, caused disruption of the nodal IXP in the region as well as all the online activities depending on it [72].

In a *distributed dynamic*, a single event disrupts simultaneously similar components scattered across a system. In this case, the systematicity is not given by a vertical chain reaction, where a disruption leads to another, but rather from structural vulnerabilities

that simultaneously affect various assets. Distributed dynamics are particularly relevant when many entities from different sectors rely on the same landscape of providers, products, and services, or in other words, share the same vulnerabilities. This trend concentrates cybersecurity risks into critical nodes, potentially magnifying the impact of events. Recently, there has been a notable surge in events that triggered distributed dynamics, highlighting how failures have the potential to escalate into systemic incidents. For instance, in November 2021, a group of researchers disclosed a critical vulnerability in the Apache Log4j software library. Log4j is a piece of open-source software which provides logging capabilities for Java applications, and that is embedded in billions of devices and systems worldwide. Exploiting this vulnerability gave the possibility to execute remote code on affected systems, leaving an open door to all sorts of malicious activities [73]. The vulnerability has triggered widespread concern and a massive effort to release patches, which is still ongoing. Further, organisations are encountering difficulties in implementing these patches. Insights from experts suggest that a complete resolution of the problem could span years, which leaves a vast number of stakeholders exposed until this issue is comprehensively addressed. Currently, there have been no reported instances of exploiting this vulnerability. However, experts agree that this can potentially trigger distributed dynamics and lead to systemic events [74].

A similar distributed dynamic led to the 2020 SolarWinds incident. SolarWinds is a software vendor which provides IT management and monitoring solutions to many clients in different industries. Hackers managed to infiltrate its software development process, injecting malicious code into their Orion platform updates. The malware was then spread across the client ecosystem as part of a legitimate software update [50]. Using SolarWinds as a vector, the malicious actors compromised more than 18,000 operators, including relevant government agencies and sensitive targets (such as the Treasury Department and Los Alamos National Laboratory, which designs nuclear weapons for the US government) as well as major ICT providers, such as Microsoft, Cisco, and FireEye [75]. While the specific details remain undisclosed, the fact that threat actors potentially accessed highly sensitive governmental information or that they could leverage the same exploit to release wipes or other destructive tools raises concern about the security around software supply chain, especially when it comes to critical operators [51].

Some authors identify a third type of systemic dynamic, the simultaneous occurrence of *independent cyber failures*. They see it as the

result of cybersecurity incidents exploiting independent vulnerabilities in single operators [76]. In theory, numerous individual cyber incidents could happen simultaneously to create a systemic event, but practically this type of event seems unlikely. For this reason, this paper focuses on top-down and distributed dynamics as the main drivers of systemic cybersecurity risk. These two scenarios are ideal types to understand how systemic cascades spread across a given environment. In concrete applications, systemic events are likely to materialise in a 'hybrid way' [1] with multiple, simultaneous, and interconnected top-down and horizontal dynamics.

In the analysed cases, the systematicity seems to stem from a confluence of factors, such as risk concentration, scale, and increased complexity of supply chains. The consolidation of cyberspace around shared assets, technologies, products, and third-party providers has created concentrated dependency on a limited set of critical nodes facilitating the establishment of shared vulnerabilities and single points of failure [77]. Moreover, the increasing complexity of computer networks and associated operational and human systems, as well as the intricate web of technical, contractual, and financial linkages on the Internet, introduces hidden levels of mutual dependence. This complexity prevents stakeholders from fully stocktaking the support that system components provide to their processes, reducing their visibility over potentially critical vulnerabilities [1].

Given the shared ownership of systemic cybersecurity risks, it is critical that all the stakeholders involved share a common understanding of the phenomenon in order to put in place meaningful and concerted mitigations. To this goal, in addition to a common definition, and analysis of systemic dynamics, practitioners need to explore new and shared approaches for identifying systemic cybersecurity risks, gaps, and vulnerabilities to enhance their capacity to address them.

7. A Review of Diagnostic Tools

It is often said that 'if you cannot measure it, you cannot manage it' [78]. Efforts to address systemic cybersecurity risks should therefore start with some sort of capacity to quantify the likelihood and severity of events as well as to identify system gaps and vulnerabilities where remediations can be applied. In this section we review some of the diagnostic tools and methodological frameworks that have been developed, and we discuss that these

efforts are undermined by a general lack of data, a partial and uneven application of methodologies as well as by a general resistance from operators to share information [79].

Given the increasing complexity, interdependency, and opacity of cyberspace, it is challenging to develop even a shared grasp of systemic cybersecurity risk, let alone efficient and consistent assessment methodologies to capture the phenomenon. Also, building this common understanding seems to be a necessary and preparatory step to develop clear regulatory frameworks for operators to manage these risks. Several efforts have been made to assess systemic cybersecurity risks. While some studies attempt to assess the individual state vulnerability to Internet infrastructure failures (such as SCC) in global comparison [67], a prevalent approach has been to leverage methods from traditional risks analysis to measure the economic impacts of cascades propagating across different linkages of a system following cybersecurity incidents. While all these studies point at the similar conclusion that direct costs associated with ‘normal’ cybersecurity incidents are significantly lower in comparison to those associated with systemic cyber events [29, 80], they also uncover concrete uncertainties in their models’ outputs. For example, a model which simulates a cybersecurity incident in a major cloud provider that disrupts service to its users estimates total losses between US\$5 and 15 billion [81]. Similarly, a recent tool to gauge the aggregated economic impact of cyber incidents in more than 60 countries through supply chain connections across various sectors estimated potential annual costs comprehended between hundreds of billions and trillions of dollars [80]. An even more emblematic example is a 2021 model to estimate the potential economic damage associated with a given cyber incident considering its cascading failures. The authors applied this model to Maersk’s NotPetya infection and found that the total economic cost may have been as little as US\$3 billion or as much as US\$57 billion [29]. These examples show significant intervals in their estimates, which in turn entail uncertainties in attempting to manage the effects on systemic cybersecurity events. The same uncertainties are highlighted in the insurance world, where catastrophe modelling is often applied to understand systemic cybersecurity risk [82], and partnerships are being endorsed to build shared datasets [83]. In fact, many authors argue that one of the main challenges that has prevented the development of approaches capable of modelling the costs and consequences of systemic failures has been the lack of data on production networks at firms’ levels, which prevents a clear understanding of interdependencies among operators [84].

A different typology of diagnostics focuses on maturity, rather than risk. These tools define a set of indicators to explore how proficiently stakeholders at different levels (from single operators to sub-sectors, sectors, and countries) implement cybersecurity measures. While these frameworks are not specifically designed to target systemic cybersecurity risks, they include measures and controls that are relevant for addressing sources of risk systematicity. More broadly, they aim to support stakeholders in building cyber robustness and resilience, which, according to recent studies, is one of the largest factors for addressing cybersecurity systemic risks [29]. Lately, most methodological frameworks have deepened their focus on systemic aspects of cybersecurity risk. At corporate levels, in 2021, the National Institute for Standards and Technology (NIST) published the 'Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations', with guidance for operators to reduce the risks associated with an enterprise's decreased visibility into and understanding of how the technology they acquire is developed, integrated, and deployed or the processes, procedures, standards, and practices used to ensure the security, resilience, reliability, safety, integrity, and quality of the products and services [85]. The Cyber Resilience Framework (CRF) and related Cyber Resilience Index (CRI) published in 2022 [86] has an even stronger focus on securing interdependencies among organisations, ecosystems, and supply chains. The CRF identifies 'systemic resilience and collaboration' as one of the six key principles that stakeholders should keep in mind while securing their assets, which entails the following 'practices': understanding the interdependencies within each ecosystem, engaging with the other relevant stakeholders and fulfilling its role in maintaining ecosystem resilience [86]. Building on the CRF, the CRI aggregates results from individual organisations and establishes an index of cyber resilience performance for sub-sectors, sectors, and supply chains. While this tool might provide a precious overview to practitioners and policymakers, its insightfulness largely depends on how broadly it is adopted by the operators forming the system as well as on the quality and accuracy of the information that is shared. This might be a significant obstacle, as organisations are often hesitant to reveal sensitive information regarding their dependencies to external parties, including government authorities. Their concerns may include the risk of losing competitive edges, attracting regulatory and legal scrutiny, or inadvertently offering a blueprint for potential adversaries to exploit. This is particularly true for large technology providers who tend to closely protect their technical architectures as a trade secret [1].

At a less granular level, the Sectoral Cybersecurity Maturity Model (SCMM) [87] and the Cybersecurity Capacity Maturity Model for

Nations (CMM) [88] aim at measuring the general cybersecurity maturity of a sector and country, respectively, and they both include relevant indicators for systemic cybersecurity risks. The SCMM builds upon the contemporary research on system science showing that an increase in resilience of individual components within a system does not necessarily result in a proportional improvement in the resilience of the system as a whole [89]. Rather, system resilience is intricately linked to the interactions among its components and is not simply the sum of the individual capacity of its constituent parts. The SCMM tries to take an approach which looks at a sector 'as a system' focusing not only on the maturity of individual components (such as critical operators) but also emphasising interdependencies and interactions among various stakeholders that constitute the sector (e.g., supervisory authorities, individual organisations, etc.) and with relevant external entities that may influence or impact the cybersecurity, capabilities, and resilience of the sector, such as Ministries, Departments, and Agencies (MDAs), national competent authorities for cybersecurity, and ICT/operational technology (OT) service providers [87]. To this end, it analyses a sector adopting, among others, indicators that look at how sector interdependencies are mapped, how information are shared among stakeholders, and how minimum levels of security are guaranteed by supply chain providers. Similarly, the CMM employs analogous indicators to assess capacity at the national level. This methodology, in addition to assessing general cybersecurity risk management and critical infrastructure protection (CIP) practices, includes specific indicators on how a country supports the resilience of Internet services and security ICT supply chain, which is particularly relevant to reducing systemic cybersecurity risks [29]. While both methodological frameworks have the potential to help countries build better security at both sectoral and national levels, including practices to target systemic cybersecurity risks, their focus on capacity, or in other words, what measures are implemented, says little about the adequacy of these measures in relation to the risk. In fact, systems are heterogeneous with different levels of digitalisation and interconnection, thus facing different risk profiles. Therefore, any capacity assessment should be contextualised and focused not primarily on what capacities are in place but rather on the process that led stakeholders to build these capacities. In particular, implementing cybersecurity measures should follow an information-driven approach. Especially for cybersecurity systemic risks, due to the increased complexity and opacity, how decision-makers identify gaps and prioritise remediation is an aspect that future research should emphasise more vigorously.

8. Conclusions

The trends and events outlined in this paper serve as a signal that systemic dynamics within cyberspace are concrete, with the potential for related risks to materialise. Nonetheless, different interpretations make it more difficult to unite stakeholders in concerted actions. Given the shared ownership of systemic cybersecurity risks, and that effective solutions demand extensive collaboration across stakeholders, establishing a common terminology and comprehension is crucial. In fact, single entities hardly have sufficient data and information, mitigations, tools, and, more broadly, capacity, to manage systemic cybersecurity risks on their own. Rather, the necessary capacity seems spread across a variety of public and private actors. Building a successful partnership among these disparate stakeholders requires not only a mutual understanding of different contextual interests and interpretations of systemic cybersecurity risk but, most importantly, a workable definition of the phenomenon itself, which, in turn, positively affects the proficiency with which stakeholders protect their assets. This is particularly relevant, as national and regional governments have started producing regulations that include requirements for operators to address systemic cybersecurity risks. For instance, the European Union Digital Operational Resilience Act (DORA) sets rules on ICT third-party risk monitoring and mitigation that highlight the need for a clearer discussion of where supply chain risk ends and where systemic cybersecurity risk begins. At the same time, the revised European Union (EU) Network and Information System Directive (NIS2) requires member states to address cybersecurity in the supply chain as part of their national cybersecurity strategies.

In this paper, we first explored existing approaches to dealing with systemic cybersecurity risk, highlighting how efforts to define and manage it result in ad hoc and uncoordinated strategies. We then proposed a comprehensive and flexible definition of systemic cybersecurity risk that could be applied at different levels of granularity, providing a common foundation for understanding and addressing the issue. Subsequently, we applied our definition to review relevant case studies, arguing that while catastrophic cybersecurity incidents are still unseen, several cybersecurity events highlight systemic dynamics. Finally, we concluded by reviewing some of the diagnostic tools and methodological frameworks that have been developed, discussing how these efforts are undermined by a general lack of data, a partial and uneven application of methodologies, and a general resistance from operators to share information.

Given the breadth and complexity of the underlying problem, new policy approaches are needed. Future research should focus on how policymakers can enhance the ability to identify and measure systemic cybersecurity risk on the one hand, and mitigate, externalise, or even eliminate it on the other. Inclusive mechanisms need to be established to involve a diversity of stakeholders: private actors, such as technology providers, cybersecurity firms, critical infrastructure operators, and reinsurers, as well as public actors, including regulators and national security agencies. International cooperation is also essential because systemic cybersecurity risk is inherently global.

References

- [1] D. Forscey, J. Bateman, N. Beecroft, B. Woods. *Systemic cyber risk: A primer*. Washington, DC: Carnegie Endowment for International Peace, 2022.
- [2] S. Romanosky, "Examining the costs and causes of cyber incidents," *Journal of Cybersecurity*, vol. 2, no. 2, pp. 121–135, 2016, doi: [10.1093/cybsec/tyw001](https://doi.org/10.1093/cybsec/tyw001).
- [3] I. Aldasoro, L. Gambacorta, P. Giudici, T. Leach, "The drivers of cyber risk," *Journal of Financial Stability*, vol. 60, pp. 100989, 2022, doi: [10.1016/j.jfs.2022.100989](https://doi.org/10.1016/j.jfs.2022.100989).
- [4] H. Labus. (Nov. 14, 2023). *Danish energy sector hit by a wave of coordinated cyberattacks* [Online]. Available: <https://www.helpnetsecurity.com/2023/11/14/danish-energy-sector-cyberattack/> [Accessed: Nov. 15, 2023].
- [5] SektorCERT. (Nov. 2023). *The attack against Danish, critical infrastructure*, SektorCERT [Online]. Available: <https://sektorcet.dk/wp-content/uploads/2023/11/SektorCERT-The-attack-against-Danish-critical-infrastructure-TLP-CLEAR.pdf> [Accessed: Feb. 12, 2024].
- [6] S.R. Das, K.J. Mitchener, A. Vossmeier. (2018). *Systemic risk and the great depression* [Online]. Available: <https://www.econstor.eu/handle/10419/198785> [Accessed: Aug. 12, 2023].
- [7] S. Eijffinger, "Defining and measuring systemic risk," in *Handbook of central banking, financial regulation and supervision: After the financial crisis*, 2011, doi: [10.4337/9781849805766.00018](https://doi.org/10.4337/9781849805766.00018).
- [8] G. Galati, R. Moessner, "Macroprudential policy – A literature review," *Journal of Economic Surveys*, vol. 27, no. 5, pp. 846–878, 2013, doi: [10.1111/j.1467-6419.2012.00729.x](https://doi.org/10.1111/j.1467-6419.2012.00729.x).
- [9] P. Smaga. (Aug. 2014). *The concept of systemic risk* [Online]. Available: <https://papers.ssrn.com/abstract=2477928> [Accessed: Aug. 12, 2023].
- [10] P. Pederson, D. Dudenhoeffer, S. Hartley, M. Permann, "Critical infrastructure interdependency modeling: A survey of US and international research," *Idaho National Laboratory*, vol. 25, pp. 27, 2006.
- [11] M. Ouyang, "Review on modeling and simulation of interdependent critical infrastructure systems," *Reliability Engineering & System Safety*, vol. 121, pp. 43–60, 2014, doi: [10.1016/j.ress.2013.06.040](https://doi.org/10.1016/j.ress.2013.06.040).

- [12] S.M. Rinaldi, J.P. Peerenboom, T.K. Kelly, "Identifying, understanding, and analyzing critical infrastructure interdependencies," *IEEE Control Systems Magazine*, vol. 21, no. 6, pp. 11–25, 2001, doi: [10.1109/37.969131](https://doi.org/10.1109/37.969131).
- [13] V. Rosato, L. Issacharoff, F. Tiriticco, S. Meloni, S. Porcellinis, R. Setola, "Modelling interdependent infrastructures using interacting dynamical models," *International Journal of Critical Infrastructures*, vol. 4, no. 1–2, pp. 63–79, 2008, doi: [10.1504/IJCIS.2008.016092](https://doi.org/10.1504/IJCIS.2008.016092).
- [14] E. Luijff, M. Klaver, "Analysis and lessons identified on critical infrastructures and dependencies from an empirical data set," *International Journal of Critical Infrastructure Protection*, vol. 35, p. 100471, 2021, doi: [10.1016/j.ijcip.2021.100471](https://doi.org/10.1016/j.ijcip.2021.100471).
- [15] D. W. Jorgenson, K.M. Vu, "The ICT revolution, world economic growth, and policy issues," *Telecommunications Policy*, vol. 40, no. 5, pp. 383–397, 2016, doi: [10.1016/j.telpol.2016.01.002](https://doi.org/10.1016/j.telpol.2016.01.002).
- [16] PCCIP. (Jan. 1998). *Critical foundations: Protecting America's infrastructures* [Online]. Available: <http://www.tandfonline.com/doi/abs/10.1080/1097198X.1998.10856225> [Accessed: Aug. 14, 2023].
- [17] A. Azadegan, M.M. Parast, L. Lucianetti, R. Nishant, J. Blackhurst, "Supply chain disruptions and business continuity: An empirical assessment," *Decision Sciences*, vol. 51, no. 1, pp. 38–73, 2020, doi: [10.1111/deci.12395](https://doi.org/10.1111/deci.12395).
- [18] G. Strupczewski, "Defining cyber risk," *Safety Science*, vol. 135, p. 105143, 2021, doi: [10.1016/j.ssci.2020.105143](https://doi.org/10.1016/j.ssci.2020.105143).
- [19] R. Böhme, S. Laube, M. Riek, "A fundamental approach to cyber risk analysis," *Variance*, vol. 12, no. 2, pp. 161–185, 2019.
- [20] E.M. Brunner and M. Suter, *International CIIP handbook 2008/2009: An inventory of 25 national and 7 international critical information infrastructure protection policies*. ETH Zürich: Center for Security Studies (CSS), 2008.
- [21] World Economic Forum (WEF). *Systemic cybersecurity risk and role of the global community: Managing the unmanageable*. Cologny, Geneva: WEF, 2022.
- [22] T. Macaulay (2019). *The danger of critical infrastructure interdependency* [Online]. Available: <https://www.cigionline.org/articles/danger-critical-infrastructure-interdependency/> [Accessed: Aug. 24, 2023].
- [23] D. Geer, E. Jardine, E. Leverett. (Feb. 2020) "On market concentration and cybersecurity risk," *Journal of Cyber Policy*, vol. 5, no. 1, pp. 9–29, doi: [10.1080/23738871.2020.1728355](https://doi.org/10.1080/23738871.2020.1728355).
- [24] M. Lehto, "Cyber-attacks against critical infrastructure," in *Cyber security*. Cham: Springer, 2022, pp. 3–42, doi: [10.1007/978-3-030-91293-2_1](https://doi.org/10.1007/978-3-030-91293-2_1).
- [25] M. Klaver, E. Luijff, "Analyzing the cyber risk in critical infrastructures," in *Issues on risk analysis for critical infrastructure protection*. IntechOpen, 2021, doi: [10.5772/intechopen.94917](https://doi.org/10.5772/intechopen.94917).
- [26] E. Luijff, M. Klaver, *Resilience approach to critical information infrastructures*. Cham: Springer, 2019, doi: [10.1007/978-3-030-05849-4](https://doi.org/10.1007/978-3-030-05849-4).
- [27] R. Setola, "How to measure the degree of interdependencies among critical infrastructures," *International Journal of System of Systems Engineering*, vol. 2, no. 1, pp. 38–59, 2010, doi: [10.1504/IJSSE.2010.035380](https://doi.org/10.1504/IJSSE.2010.035380).

- [28] D. Clemente, *Cyber security and global interdependence: what is critical?* London: Chatham House, Royal Institute of International Affairs, 2013.
- [29] J.W. Welburn, A.M. Strong, "Systemic cyber risk and aggregate impacts," *Risk Analysis*, vol. 42, no. 8, pp. 1606–1622, 2022, doi: [10.1111/risa.13715](https://doi.org/10.1111/risa.13715).
- [30] D.J. Bodeau, C.D. McCollum. (2018) *System-of-systems threat model* [Online]. Available: <https://apps.dtic.mil/sti/citations/AD1108059> [Accessed: Aug. 15, 2023].
- [31] Office of Financial Research (OFR). "Cybersecurity and financial stability: Risks and resilience," *OFR Viewpoint*, 17-01, Feb. 2017.
- [32] P. Sommer, I. Brown, *Reducing systemic cybersecurity risk*. Organisation for Economic Cooperation and Development (OECD), Working Paper No. IFP/WKP/FGS, vol. 3. Paris: OECD, 2011.
- [33] European Systemic Risk Board (ESRB). (2020). *Systemic cyber risk* [Online]. Available: <https://data.europa.eu/doi/10.2849/566567> [Accessed: Aug. 15, 2023].
- [34] J. Fell, N. de Vette, S. Gardó, B. Klaus, J. Wendelborn. (Nov 2022). "Towards a framework for assessing systemic cyber risk," *Financial Stability Review* [Online]. Available: https://www.ecb.europa.eu/pub/financial-stability/fsr/special/html/ecb.fsrart202211_03-9a8452e67a.en.html [Accessed: Aug. 18, 2023].
- [35] A. Masys, "Examining systemic risk in the cyber landscape," in *The great power competition*, vol. 3, *Cyberspace: The fifth domain*. Cham: Springer, pp. 69–82, 2022, doi: [10.1007/978-3-031-04586-8_4](https://doi.org/10.1007/978-3-031-04586-8_4).
- [36] World Economic Forum (WEF). *Understanding systemic cyber risk*. White Paper. Cologny, Geneva: WEF, Oct. 2016
- [37] R.D. Arnold, J.P. Wade, "A definition of systems thinking: A systems approach," *Procedia Computer Science*, vol. 44, pp. 669–678, 2015, doi: [10.1016/j.procs.2015.03.050](https://doi.org/10.1016/j.procs.2015.03.050).
- [38] A. Kossiakoff, S.M. Biemer, S.J. Seymour, D.A. Flanagan, *Systems engineering principles and practice*. Hoboken, NJ: John Wiley, 2020, doi: [10.1002/9781119516699](https://doi.org/10.1002/9781119516699).
- [39] US Cybersecurity and Infrastructure Security Agency (CISA). (2021). *Systemic cyber risk reduction venture* [Online]. Available: https://www.cisa.gov/sites/default/files/2023-02/fs_systemic-cyber-risk-reduction_508.pdf [Accessed Feb. 12, 2024].
- [40] White House. (Mar. 2023). *National cybersecurity strategy* [Online]. Available: <https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Cybersecurity-Strategy-2023.pdf> [Accessed Feb. 12, 2024].
- [41] Digital Director Network (DDN). (Dec. 08, 2019). *DDN releases DiRECTOR the only systemic risk framework focused on complex digital systems*. [Online]. Available: <https://www.digitaldirectors.network:443/blogs/ddn-releases-director-the-only-systemic-risk-framework-focused-on-complex-digital-systems> [Accessed: Dec. 08, 2023].
- [42] B. Zukis. (Dec. 09, 2019). *Digital directors network releases DiRECTORTM the only systemic risk framework focused on complex digital systems* [Online]. Available: <https://www.einpresswire.com/article/504281736/digital-directors-network-releases-director-the-only-systemic-risk-framework-focused-on-complex-digital-systems> [Accessed: Aug. 21, 2023].

- [43] I. Smith. (Aug. 18, 2023). "Cyber attacks set to become 'uninsurable', says Zurich chief," *Financial Times* [Online]. Available: <https://www.ft.com/content/63ea94fa-c6fc-449f-b2b8-ea29cc83637d> [Accessed: Dec. 26, 2026].
- [44] AIG. (2017). *Is cyber risk systemic* [Online]. Available: https://insidecybersecurity.com/sites/insidecybersecurity.com/files/documents/may2017/cs2017_0167.pdf [Accessed Feb. 12, 2024]
- [45] A. Granato, A. Polacek. (2019). *The growth and challenges of cyber insurance – Federal Reserve Bank of Chicago*, Chicago Fed Letter No. 426 [Online]. Available: <https://www.chicagofed.org/publications/chicago-fed-letter/2019/426> [Accessed: Aug. 18, 2023], doi: [10.21033/cfl-2019-426](https://doi.org/10.21033/cfl-2019-426).
- [46] S. Scalfane. (2021). *Writing cyber is key to survival, Munich Re Exec says* [Online]. Available: <https://www.carriermanagement.com/news/2021/09/13/226172.htm> [Accessed: Aug. 18, 2023].
- [47] I. Smith, "Lloyd's of London defends cyber insurance exclusion for state-backed attacks," *Financial Times*, Sep. 05, 2022.
- [48] T. Rid, B. Buchanan, "Attributing cyber attacks," *Journal of Strategic Studies*, vol. 38, no. 1–2, pp. 4–37, 2015, doi: [10.1080/01402390.2014.977382](https://doi.org/10.1080/01402390.2014.977382).
- [49] M. Mueller, K. Grindal, B. Kuerbis, F. Badiei, "Cyber attribution," *Cyber Defense Review*, vol. 4, no. 1, pp. 107–122, 2019.
- [50] R. Alkhadra, J. Abuzaid, M. AlShammari, N. Mohammad. (2021). "Solar winds hack: In-depth analysis and countermeasures," in *2021 12th International conference on computing communication and networking technologies (ICCCNT)*, IEEE, pp. 1–7, doi: [10.1109/ICCCNT51525.2021.9579611](https://doi.org/10.1109/ICCCNT51525.2021.9579611).
- [51] W. Growley, L. Gruden, W. Canter, "Navigating the solar winds supply chain attack," vol. 56, no. 2, 2021.
- [52] EWI. (2019). *Cyber insurance and systemic market risk*. EastWest Institute [Online]. Available: <https://www.eastwest.ngo/cyberinsurance> [Accessed: Aug. 16, 2023].
- [53] S. Smith, *Out of gas: A deep dive into the colonial pipeline cyberattack*. SAGE Business Cases Originals. Thousand Oaks, CA: SAGE, 2022, doi: [10.4135/9781529605679](https://doi.org/10.4135/9781529605679).
- [54] J.W. Goodell, S. Corbet, "Commodity market exposure to energy-firm distress: Evidence from the Colonial pipeline ransomware attack," *Finance Research Letters*, vol. 51, pp. 103329, 2023, doi: [10.1016/j.frl.2022.103329](https://doi.org/10.1016/j.frl.2022.103329).
- [55] S. Ghafur, S. Kristensen, K. Honeyford, G. Martin, A. Darzi, P. Aylin, "A retrospective impact analysis of the WannaCry cyberattack on the NHS," *NPJ Digital Medicine*, vol. 2, no. 1, pp. 1–7, 2019, doi: [10.1038/s41746-019-0161-6](https://doi.org/10.1038/s41746-019-0161-6).
- [56] White House. (2018). *Statement from the Press Secretary – The White House* [Online]. Available: <https://trumpwhitehouse.archives.gov/briefings-statements/statement-press-secretary-25/> [Accessed: Aug. 17, 2023].
- [57] C. Nast (Aug. 21, 2018). *The untold story of NotPetya, the most devastating cyber-attack in history* [Online]. Available: <https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/> [Accessed: Aug. 17, 2023].
- [58] A. Jones, O. Khan, "Surviving NotPetya: Global supply chains in the era of the cyber weapon," in *Cyber security and supply chain management: Risks, challenges, and solutions*, pp. 133–146, 2021, doi: [10.1142/9789811233128_0006](https://doi.org/10.1142/9789811233128_0006).

- [59] A. Satariano, N. Perlroth (Apr. 15, 2019). "Big companies thought insurance covered a cyberattack. They may be wrong," *The New York Times* [Online]. Available: <https://www.nytimes.com/2019/04/15/technology/cyberinsurance-notpetya-attack.html> [Accessed: Aug. 18, 2023].
- [60] ISO. (2016). *ISO guide 73: 2009* [Online]. Available: <https://www.iso.org/standard/44651.html> [Accessed: Aug. 22, 2023].
- [61] K. Stine, R. Kissel, W.C. Barker, J. Fahlsing, J. Gulick, *Guide for mapping types of information and information systems to security categories*, vol. 1. 2008, doi: [10.6028/NIST.SP.800-60v1r1](https://doi.org/10.6028/NIST.SP.800-60v1r1).
- [62] M.H. Hugos. *Essentials of supply chain management*. New York, NY: John Wiley, 2024.
- [63] F. Rieger. (Jan. 17, 2019). *Stuxnet: targeting the Iranian enrichment centrifuges in Natanz? Knowledge brings fear* [Online]. Available: <https://frank.geekheim.de/?p=1189> [Accessed: Nov. 08, 2023].
- [64] J.R. Lindsay, "Stuxnet and the limits of cyber warfare," *Security Studies*, vol. 22, no. 3, pp. 365–404, 2013, doi: [10.1080/09636412.2013.816122](https://doi.org/10.1080/09636412.2013.816122).
- [65] NATO CCDCOE. (2019). *Strategic importance of, and dependence on, undersea cables* [Online]. Available: <https://ccdcoe.org/uploads/2019/11/Undersea-cables-Final-NOV-2019.pdf> [Accessed: Aug. 22, 2023].
- [66] ESCAP, "Broadband Connectivity in Pacific Island Countries," 2018
- [67] J. Franken, T. Reinhold, L. Reichert, C. Reuter, "The digital divide in state vulnerability to submarine communications cable failure," *International Journal of Critical Infrastructure Protection*, vol. 38, p. 100522, 2022, doi: [10.1016/j.ijcip.2022.100522](https://doi.org/10.1016/j.ijcip.2022.100522).
- [68] G. Aceto, A. Botta, P. Marchetta, V. Persico, A. Pescapé, "A comprehensive survey on Internet outages," *Journal of Network and Computer Applications*, vol. 113, pp. 36–63, 2018, doi: [10.1016/j.jnca.2018.03.026](https://doi.org/10.1016/j.jnca.2018.03.026).
- [69] G.M. Graff. (Dec. 13, 2017). *How a dorm room "Minecraft" scam brought down the Internet* [Online]. Available: <https://www.wired.com/story/mirai-botnet-minecraft-scam-brought-down-the-internet/> [Accessed: Aug. 23, 2023].
- [70] L.H. Newman. (Oct. 21, 2016). *What we know about Friday's massive East Coast Internet outage* [Online]. Available: <https://www.wired.com/2016/10/internet-outage-ddos-dns-dyn/> [Accessed: Aug. 23, 2023].
- [71] L. Chadwick. (Jul. 22, 2019). *Italy hit by four-hour train delays after suspected arson fire* [Online]. Available: <https://www.euronews.com/2019/07/22/italy-hit-by-four-hour-train-delays-after-suspected-arson-fire-outside-florence> [Accessed: Aug. 23, 2023].
- [72] M. Faye. (Mar. 14, 2023). *When mother nature strikes: Lessons learned from an IXP crash* [Online]. Available: <https://www.linkedin.com/pulse/when-mother-nature-strikes-lessons-learned-from-ixp-crash-faye> [Accessed: Aug. 23, 2023].
- [73] P. Ferreira, F. Caldeira, P. Martins, M. Abbasi, "Log4j vulnerability," in: *Information technology and systems*. Cham: Springer, 2023, pp. 375–385, doi: [10.1007/978-3-031-33261-6_32](https://doi.org/10.1007/978-3-031-33261-6_32).

- [74] J. Marks. (Jan. 11, 2022). *Analysis: One month in, there aren't any huge, known log4j hacks* [Online]. Available: <https://www.washingtonpost.com/politics/2022/01/11/one-month-there-arent-any-huge-log4j-hacks/> [Accessed: Aug. 25, 2023].
- [75] K. Zetter. (May 02, 2023). *The untold story of the boldest supply-chain hack ever* [Online]. Available: <https://www.wired.com/story/the-untold-story-of-solarwinds-the-boldest-supply-chain-hack-ever/> [Accessed: Nov. 15, 2023].
- [76] E.D. Peet, M.J. Vermeer. (2020). *Securing communications in the quantum computing age: Managing the risks to encryption* [Online]. Available: <https://policycommons.net/artifacts/4835890/securing-communications-in-the-quantum-computing-age/5672600/> [Accessed: Nov. 11, 2023].
- [77] J.E. Scheuermann, "Cyber risks, systemic risks, and cyber insurance symposium," *Penn State Law Review*, vol. 122, no. 3, pp. 613–644, 2017.
- [78] J. Drucker. (Dec. 04, 2018). *Council post: You are what you measure* [Online]. Available: <https://www.forbes.com/sites/theyec/2018/12/04/you-are-what-you-measure/> [Accessed: Nov. 10, 2023].
- [79] F. Cremer et al. "Cyber risk and cybersecurity: A systematic review of data availability," *The Geneva Papers on Risk and Insurance – Issues and Practice*, vol. 47, no. 3, pp. 698–736, 2022, doi: [10.1057/s41288-022-00266-6](https://doi.org/10.1057/s41288-022-00266-6).
- [80] P. Dreyer et al. (2018). *Estimating the global cost of cyber risk*. Research Reports RR-2299-WFHF. Rand Corporation [Online]. Available: https://www.rand.org/content/dam/rand/pubs/research_reports/RR2200/RR2299/RAND_RR2299.pdf [Accessed: Nov. 11, 2023].
- [81] Lloyd's. (2018). *Cloud down impacts on the US economy* [Online]. Available: <https://assets.lloyds.com/assets/pdf-air-cyber-lloyds-public-2018-final/1/pdf-air-cyber-lloyds-public-2018-final.pdf> [Accessed: Aug. 22, 2023].
- [82] Gallagher Re. (Jan. 2019). *Evaluation of cyber models* [Online]. Available: <https://www.ajg.com/gallagherre/-/media/files/gallagher/gallagherre/evaluating-cyber-models-report.pdf> [Accessed: Aug. 22, 2023].
- [83] C. Shi. (Sep. 28, 2022). *CFC spearheads cyber cat-declaration initiative to tackle systemic risk* [Online]. Available: <https://www.insuranceinsider.com/article/2aoh41qs3atr6x3jpj75s/reinsurers-section/cfc-spearheads-cyber-cat-declaration-initiative-to-tackle-systemic-risk> [Accessed: Apr. 18, 2024].
- [84] V.M. Carvalho, A. Tahbaz-Salehi, "Production networks: A primer," *Annual Review of Economics*, vol. 11, no. 1, pp. 635–663, 2019, doi: [10.1146/annurev-economics-080218-030212](https://doi.org/10.1146/annurev-economics-080218-030212).
- [85] J. Boyens, A. Smith, N. Bartol, K. Winkler, A. Holbrook, et al., "Cybersecurity supply chain risk management practices for systems and organizations," Oct. 2021, doi: [10.6028/NIST.SP.800-161r1-draft2](https://doi.org/10.6028/NIST.SP.800-161r1-draft2).
- [86] World Economic Forum (WEF) and Accenture. (Jul. 2022). *The cyber resilience index: Advancing organizational cyber resilience* [Online]. Available: <https://www.weforum.org/publications/the-cyber-resilience-index-advancing-organizational-cyber-resilience/> [Accessed: Nov. 11, 2023].
- [87] The World Bank Group. (2023). *Sectoral cybersecurity maturity model* [Online]. Available: <https://documents1.worldbank.org/curated/en/099062623085028392/pdf/P17263707c36b702309f7303dbb7266e1cf.pdf> [Accessed: Feb. 12, 2024].

- [88] *Global Cyber Security Capacity Centre (GCSCC)*. (2021). *Cybersecurity capacity maturity model for nations* [Online]. Available: <https://gcsc.ox.ac.uk/the-cmm> [Accessed: Nov. 12, 2023], doi: [10.2139/ssrn.3822153](https://doi.org/10.2139/ssrn.3822153).
- [89] L. D. Valdez et al. "Cascading failures in complex networks," *Journal of Complex Networks*, vol. 8, no. 2, p. cnaa013, 2020, doi: [10.1093/comnet/cnaa013](https://doi.org/10.1093/comnet/cnaa013).

Enhancing Secure Key Management Techniques for Optimised 5G Network Slicing Security

Kovid Tiwari | School of Computing Science Engineering & Artificial Intelligence, VIT Bhopal University, India | ORCID: 0009-0007-5838-7593

Ajay Kumar Phulre | School of Computing Science Engineering & Artificial Intelligence, VIT Bhopal University, India | ORCID: 0000-0001-7457-1007

Devraj Vishnu | School of Computing Science Engineering & Artificial Intelligence, VIT Bhopal University, India | ORCID: 0000-0002-3106-2939

Saravanan D | School of Computing Science Engineering & Artificial Intelligence, VIT Bhopal University, India | ORCID: 0000-0001-8992-6755

Abstract

This research enhances the security of 5G network slicing by introducing a Secure Key Management (SKM) framework designed to protect data within virtualised network environments. Network slicing, while a transformative feature of 5G, introduces complex vulnerabilities, especially intra-slice and inter-slice threats, which require specialised security mechanisms. This study addresses these risks by proposing a mathematically-driven SKM model that combines Shamir's Secret Sharing (SSS) and homomorphic encryption for secure key generation and distribution. The model guarantees that threats of unauthorised access are reduced to a minimum while maintaining efficiency within the contexts of a multi-slice environment. One of the major contributions presented in this paper is proposing a correlation engine that is implemented as a part of the SKM framework for real-time detection of inter-slice as well as intra-slice attacks. In order to prove the efficiency of the used framework, it was applied in the experimental 5G slicing setup

Received: 14.11.2024

Accepted: 29.12.2024

Published: 31.12.2024

Cite this article as:

K. Tiwari, A.K. Phulre, D. Vishnu, D. Saravanan, "Enhancing secure key management techniques for optimised 5G network slicing security," ACIG, vol. 3, no. 2, 2024, pp. 170–210. DOI: 10.60097/ACIG/200243

Corresponding author:

Kovid Tiwari, School of Computing Science Engineering & Artificial Intelligence, VIT Bhopal University, India; E-mail: kovidtiwarifeb3@gmail.com

 0009-0007-5838-7593

Copyright:

Some rights reserved (CC-BY):

Kovid Tiwari, et al.
Publisher NASK



under various attack conditions. From the results the benefit of the proposed methods was identified which include the reduction of data leakage risks and lower Denial of Service (DoS) compared to the baseline. Notably, the proposed model enhanced the efficiency of the slice isolation and key distribution while at the same time strengthening its security and performance. In an attempt to combine theoretical models with practical validation, this research will offer a holistic security model for 5G network slicing that directly solves scalability and dynamic key management. The results enrich the literature on security enhancement for next-generation telecommunication networks and provide a strong basis for real-world experimentation.

Keywords

5G slicing, secure key management, intra-slice security, network security, cryptographic models

1. Introduction

The rapid expansion of 5G technology has transformed telecommunications by offering faster speeds, ultra-low latency and the ability to support diverse services, from enhanced mobile broadband (eMBB) to massive Machine Type Communications (mMTC) [1]. The other significant transformation of the 5G architecture is network slicing, defined as the ability to create multiple logical networks on the same physical infrastructure to meet different use cases. Complementing the latter is the fact that network slicing presents unusual security threats in terms of the slices' identity, confidentiality and accessibility. Security threats in 5G slicing can be broadly classified into intra-slice attacks and inter-slice attacks. Intra-slice attacks occur when an adversary exploits a vulnerability within a single slice, potentially compromising sensitive data or disrupting services.. To address these challenges, this research proposes a Secure Key Management (SKM) framework specifically designed for 5G network slicing security. The SKM system integrates Shamir's Secret Sharing (SSS) for secure key generation and homomorphic encryption for confidential data handling within network slices [4]. These techniques guarantee that the keys will be in possession of different parties to avoid centralisation and hence minimise cases of compromise. Moreover, a correlation engine is incorporated for inter-slice and intra-slice anomaly detect and defence mechanisms operatively. These slices correlate the behaviour of the network and detect anomalies and suspicious accesses to improve the predictive abilities of threat detection.

The contribution of this research is that, for the first time, it employs mathematical models to provably establish and optimise secure key management tailored to 5G slicing dynamics. While with the traditional models, the use of keys is quite fixed through key distribution and the isolation between slices is often less effective, the proposed new SKM system is efficient in dynamic traffic loads and security requirements for individual slices [7]. In addition, the homomorphic encryption framework improves data security while reducing the effects on system response time, while the correlation engine continuously counteracts threats. In addition to presenting a conceptual security framework, this work demonstrates the usefulness of this framework by verifying the proposed security models through an emulated 5G network slicing platform environment. The evaluations based on our experiments are as follows: While compared with the baseline methods, the framework can effectively prevent the Denial of Service (DoS) attacks, decrease the data leakage risks and improve the slice isolation.

By bridging theoretical concepts with practical implementation, this research makes a significant contribution to the 5G security landscape [8]. It offers a scalable and efficient solution for intra-slice and inter-slice security, positioning SKM as a vital component in the evolving 5G architecture.

2. Literature Review

Network slicing is an essential technique in 5G networks, but it introduces new security challenges. Two key approaches to addressing these challenges are isolation and secure key management. Isolation methods are designed to prevent attacks from spreading across different slices within the network. This can be achieved using tools such as firewalls, VLANs and network function virtualisation (NFV). On the other hand, SKM plays a crucial role in safeguarding the data and traffic within network slices, as keys are essential for encrypting data and authenticating both users and devices. While there has been significant research on key management for network slicing, several challenges remain [11]. One major issue is the development of scalable and efficient key management systems, as many existing methods are tailored for traditional networks and may not be suited to the dynamic, complex nature of network slicing. Additionally, improved support for inter-domain and inter-operator key management is needed, as 5G network slices often span multiple domains, including the radio access network (RAN), core network and cloud environments. Network slices may need to span across numerous operators. This project intends to

build new and novel key management techniques to aid secure network slicing. The suggested vital management techniques will be designed to be efficient, scalable and safe against physical threats. The suggested important management techniques will also provide inter-domain and inter-operator key management. This paper provides a complete overview of 5G network slicing security aspects. It goes into potential dangers and responses, presenting insights into preserving the varied slices from inter and intra-slice attacks [14]. The conclusion underscores the importance of a holistic security approach in the 5G network slicing paradigm. This study focuses on the scalability and flexibility of network slicing in 5G and examines its impact on enhancing network efficiency while addressing security challenges. It highlights the need for adaptive security strategies to match the dynamic nature of sliced networks. One key recommendation is the use of slice isolation to mitigate Distributed Denial of Service (DDoS) attacks targeting 5G core network slices. The findings stress the importance of implementing isolation techniques to bolster the resilience of individual slices against such attacks, contributing to a more robust and secure 5G infrastructure. In addition, the study delves into the security and privacy-preserving aspects of network slicing within the 3GPP 5G architecture [12]. It emphasises the critical need for robust privacy protections to safeguard sensitive data and ensure user privacy, ultimately promoting a secure and trusted 5G environment.

This framework specifies criteria for implementing network slicing in 5G. It covers the outcome of adopting a standardised approach to network slicing, providing a framework for secure and interoperable implementations throughout the 5G ecosystem. The study offers the VIKOR technique for efficient and secure 5G core network slice provisioning [15]. The outcome illustrates the usefulness of this method in improving resource allocation while ensuring the security of specific network slices in the 5G core. Focused on key management, this study provides a safe keying strategy for network slicing in 5G. Although significant strides have been made in securing 5G network slicing using techniques such as encryption, VLAN tagging and blockchain, major security challenges remain. These include inadequate dynamic slice isolation, inefficient key management and the lack of real-time threat detection. This study addresses these gaps with the introduction of a SKM framework, incorporating SSS, homomorphic encryption and a real-time correlation engine. The proposed model not only enhances the security of 5G network slicing but also provides a scalable solution that can adapt to the dynamic nature of 5G environments [17]. Table 1 provides an overview of various research efforts in the field of 5G

Table 1. Analytical representation of ‘Top Researches Papers’.

Papers Name	Year	Authors	Field of Research	Challenges	Results	Effectiveness in %
Network Slicing Scalability Flexibility in 5G Networks [20]	2021	3GPP	3GPP Network Design	Scalability, Resource Isolation	Introduced Network Slicing concept and basic principles	The revenue potential of network slicing is 82%
Towards Secure Slicing: Using Slice Isolation to Mitigate DDoS Attacks on 5G Core Network Slices [5]	2020	Y. Zhang, Y. Liu, and Y. Li	5G Core Network Security	DDoS attacks, slice isolation	Proposed isolation techniques for DDoS mitigation within slices	85% reduction in DDoS attack impact
Secure and Privacy-preserving Network Slicing in 3GPP 5G System Architecture [6]	2019	Y. Liu, Y. Zhang, Zhang et al.	Privacy-Preserving Slicing	Privacy Leakage, Slice Isolation, Identity Management	Develop a privacy-preserving mechanism for slicing while ensuring isolation	98% reduction in privacy leakage risk
161010_NGMN Network Slicing framework v1.0.8	2018	NGMN Alliance	Network Slicing Framework	Architecture, taxonomy, challenges, security issues, attacks classification, possible solutions, future scope	Defined a reference framework for multi-operator network slicing	Techniques such as resource isolation, cryptography and machine learning
Efficient and Secure 5G Core Network Slice Provisioning Based on VIKOR Approach [15]	2021	X. Wang, Y. Zhang, and Y. Li	Network Slicing Optimisation	Architecture, taxonomy, challenges, security issues, attacks classification, possible solutions, future scope	Defined a reference framework for multi-operator network slicing	Techniques such as resource isolation, cryptography and machine learning
Secure Keying Scheme for Network Slicing in 5G Architecture	2020	S. Kim, J. Lee, and J. Kim	Secure Key Management	Proposed a secure keying scheme for network slicing in 5G architecture	Developed a secure keying scheme for dynamic slice creation and isolation	99% success rate in crucial distribution, negligible leakage risk
Network Slicing in 5G: Survey and Challenges [12]	2020	M. A. Imran, and M. A. Qadir	Network Slicing Overview	Security, isolation, resource sharing, performance, QoS	Identified key security challenges and opportunities in network slicing	A valuable resource for researchers and practitioners interested in networking
The Isolation Concept in the 5G Network Slicing [25]	2021	Chen et al.	Slice Isolation Mechanism	Proposed an isolation-based approach to enhance security in the network slicing	Analysed different isolation techniques and their effectiveness	Provider Management, Tenant Management and the Means of Isolation

(continues)

Table 1. Analytical representation of ‘Top Researches Papers’.

Papers Name	Year	Authors	Field of Research	Challenges	Results	Effectiveness in %
Secure5G: A Deep Learning Framework Towards a Secure Network Slicing in 5G and Beyond [26]	2022	Wang et al.	Deep Learning for Security	Developed a deep learning-based framework to detect and eliminate threats in the 5G network slicing	Developed a deep learning framework for securing network slicing	92% accuracy in anomaly detection, 85% reduction in intrusion attempts
Network Slicing for 5G: Challenges and Opportunities [27]	2021	Boulogne et al.	Network Slicing Applications	Architecture, taxonomy, challenges security issues, attacks classification, possible solutions, future scope	Explored potential applications and challenges of network slicing in various scenarios	77% Network slicing in 5G, such as security, isolation, resource sharing, performance, and QoS
Towards Secure and Intelligent Network Slicing for 5G Networks [18]	2018	Asan et al.	Intelligent Slicing Security	Security, machine learning, cryptography	Proposed a framework for secure and intelligent network slicing with trust manage menu	95% reduction in security incidents, 80% improvement in automation efficiency
A Survey of Mobility Management as a Service in Real-Time Inter/Intra Slice Control [11]	2019	Kim et al.	Mobility Management in Slicing	Mobility management, service control	Surveyed existing solutions for mobility management in multi-slice scenarios	69% Evaluation of the maturity of current proposals
Classification of network slicing threats based on slicing enablers: A survey [19]	2019	Abedin et al.	Threat Analysis in Slicing	Threat classification, slicing enablers	Categorised and analysed threats to network slicing based on different slicing enablers	Threats based on slicing enablers

network slicing, covering topics like scalability, security, resource isolation and privacy. It includes studies on network slicing frameworks, isolation techniques, key management schemes and deep learning-based security methods, presenting results such as reductions in attack impact and privacy leakage, and improvements in automation efficiency and threat detection accuracy.

2.1. Overview

Network slicing in 5G has emerged as a transformative solution to meet the diverse and demanding service requirements of various applications. The core idea behind 5G network slicing is to create multiple virtual networks, or 'slices', each designed to optimise specific use cases such as enhanced Mobile Broadband (eMBB), Ultra-Reliable Low Latency Communications (URLLC) and massive Machine Type Communications (mMTC). This shift towards tailored network architectures is crucial for addressing the distinct needs of different applications in the 5G ecosystem. By leveraging a single physical network, operators can create isolated logical networks, each with customised capabilities, thus enhancing both flexibility and efficiency. At the heart of network slicing lies the ability to allocate resources dynamically across different slices to cater to the diverse needs of users and devices [4]. A key challenge in implementing 5G network slicing is ensuring secure communication within each slice, especially as the slices are isolated yet interconnected. This isolation is vital to prevent unauthorised access and ensure that critical applications such as eMBB or URLLC are protected from potential attacks. Additionally, slicing introduces a need for fine-grained quality of service (QoS) guarantees, allowing different slices to receive the appropriate levels of latency, throughput and reliability based on their specific application needs. Despite the promise of network slicing, security remains a major concern, especially in multi-slice environments where interactions between slices could potentially expose the network to attacks. One of the critical areas of focus is secure key management, which plays an essential role in safeguarding slice communications [13]. Existing key management approaches often lack the flexibility required to adapt to dynamic slice configurations. Many traditional methods rely on static key distribution mechanisms, which are inadequate for handling the dynamic nature of 5G network slicing, where slices are created, modified and terminated based on real-time needs. Recent advancements have attempted to address this issue through the integration of cryptographic techniques such as SSS and homomorphic encryption, which aim to provide decentralised and secure methods for key distribution and data confidentiality.

However, while these methods offer improvements in key security and slice isolation, they still leave gaps in real-time threat detection and response mechanisms. For example, many of the existing frameworks focus heavily on encryption and key distribution but fail to incorporate proactive measures for monitoring intra-slice and inter-slice activities to identify and prevent unauthorised access or attacks [30]. To fill these gaps, the proposed SKM framework offers a novel approach by combining SSS, homomorphic encryption and a real-time correlation engine for threat detection. The SKM framework ensures that key distribution is decentralised and adaptable, addressing the limitations of traditional centralised systems. By incorporating a dynamic threat detection mechanism, the SKM framework also provides proactive protection against both intra-slice and inter-slice attacks, offering a multi-layered defence that enhances the overall security of 5G network slices. In addition to the improved key management, the SKM framework offers a detailed mathematical model that ensures both security and efficiency in the management of keys across dynamic slices [28]. Unlike existing models that only describe encryption methods, this approach introduces concrete equations for key generation, distribution and threat mitigation, providing a solid foundation for practical implementation. Figure 1 represents the signalling flow in a 5G network slicing architecture, focusing on slice selection, PDU session setup and traffic mapping for secure communication management. While network slicing enables greater flexibility and resource optimisation, its security challenges require innovative solutions that adapt to the dynamic nature of 5G environments. This paper contributes by introducing a comprehensive SKM framework that not only secures key management but also provides a robust, scalable and efficient solution to protect the integrity of 5G network slices. The future of 5G network security depends on integrating such advanced models that address both the technical and operational challenges of network slicing, ensuring that security evolves alongside network capabilities.

2.2. Design Challenges

Network slicing, a new notion in the landscape of 5G networks, introduces the possibility of building several virtual networks atop a shared physical infrastructure. This paradigm change is crucial in enabling exceptional flexibility, scalability and customisation to satisfy the unique requirements of various applications and services. However, the fulfilment of network slicing's great promise is accompanied by a spectrum of design obstacles that span technological intricacies, architectural concerns and operational nuances. Firstly,

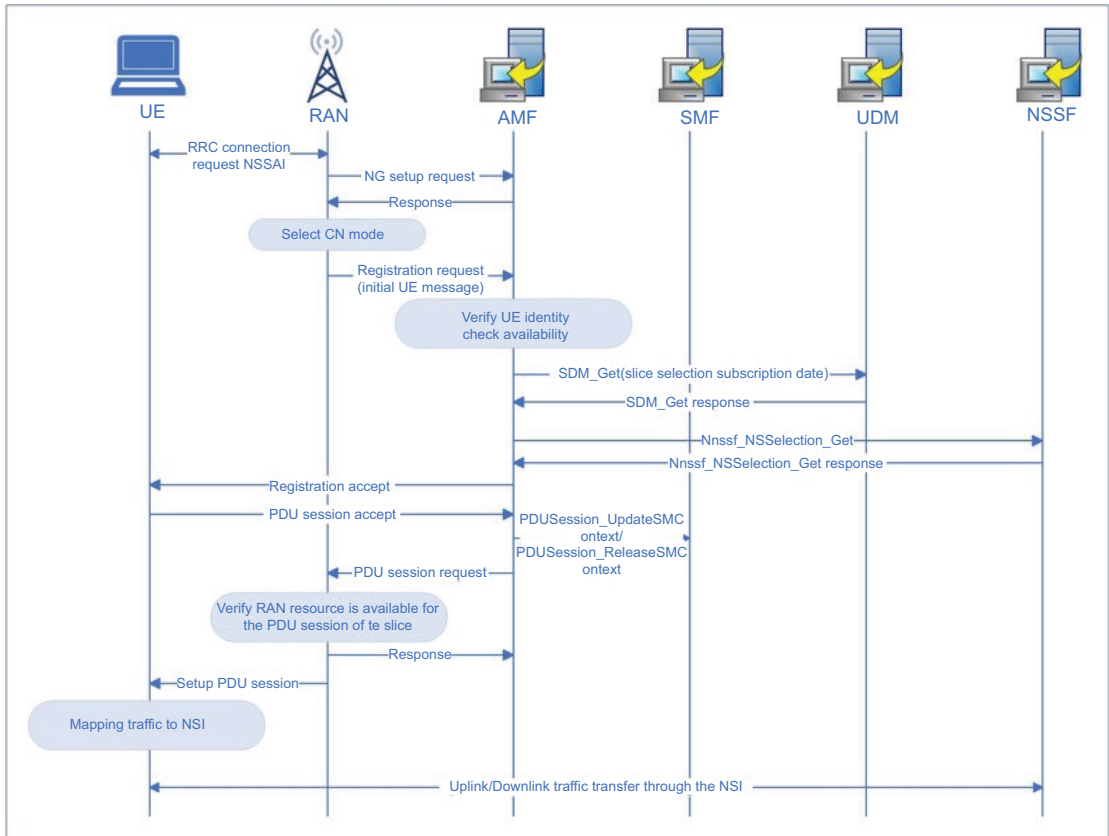


Figure 1. Network slicing processing.

the granularity limits in spectrum and radio-level resource sharing constitute a substantial difficulty [33]. Unlike fixed network slices that may be easily expanded with extra hardware resources, RAN slicing confronts a physical barrier due to the constrained availability of spectrum. Achieving this separation requires robust systems for resource allocation, bandwidth management and interference reduction. The problem lies in establishing algorithms and protocols that dynamically distribute resources depending on the variable demands of distinct slices, enhancing the overall network efficiency.

- **Orchestration and Management:** The orchestration of network slices involves coordinating and managing various resources within and across slices. When you're making a complete orchestration system, you have to think about things like enforcing policies, monitoring in real-time, and making decisions automatically [28]. Achieving the optimal equilibrium between centralised and distributed orchestration poses a significant challenge.

- **Latency and Quality of Service (QoS):** Applications have diverse latency and QoS requirements. Network slicing aims to cater to these needs, but achieving ultra-low latency and high QoS across slices is a significant challenge. Additionally, latency requirements for emerging technologies like augmented reality and autonomous vehicles pose unique hurdles.
- **Security and Privacy Concerns:** It is critical to give assurance of the privacy and security of each network slice, which creates numerous logical networks on a shared infrastructure. Creating resilient security procedures to deter illegal access, data breaches and attacks on individual slices is a complex task [23]. The architecture should consider authentication, authorisation, and encryption methods that may be adjusted to meet the unique requirements of each slice while yet ensuring a consistent security foundation for the entire network.
- **Inter-Slice Interactions:** Network slices are not isolated islands; they frequently need to interact with each other to provide end-to-end services. A significant design challenge is ensuring seamless communication and coordination between slices without compromising their independence. Inter-slice interactions involve addressing signalling, data exchange, cross-slice resource coordination issues, and standardising protocols and interfaces for inter-slice communication.

2.3. Architecture

The general design of network slicing has three levels, each with its own management functions. Figure 2 presents a network slicing architecture for mobile networks, illustrating the end-to-end service management and orchestration across different layers, including RAN, core network, transport and cloud management, enabling slice-specific functionalities.

- **Resource Layer (RL):** The foundational layer consists of network resources and functions that provide services to end-users upon request. These resources, whether physical or virtual, include storage, processing power and transmission nodes, while network functions cover routing, switching, slice selection and authentication processes.
- **Network Slice Instance Layer (NIL):** The middle layer comprises network slices, each delivering the specific capabilities needed by service instances [6]. A slice can operate directly on network resources or on another slice, supporting one or multiple service instances. Different slices may or may not share the same physical infrastructure and network functionalities.
- **Service Instance Layer:** The upper layer consists of service instances that utilise the network slices and deliver them to

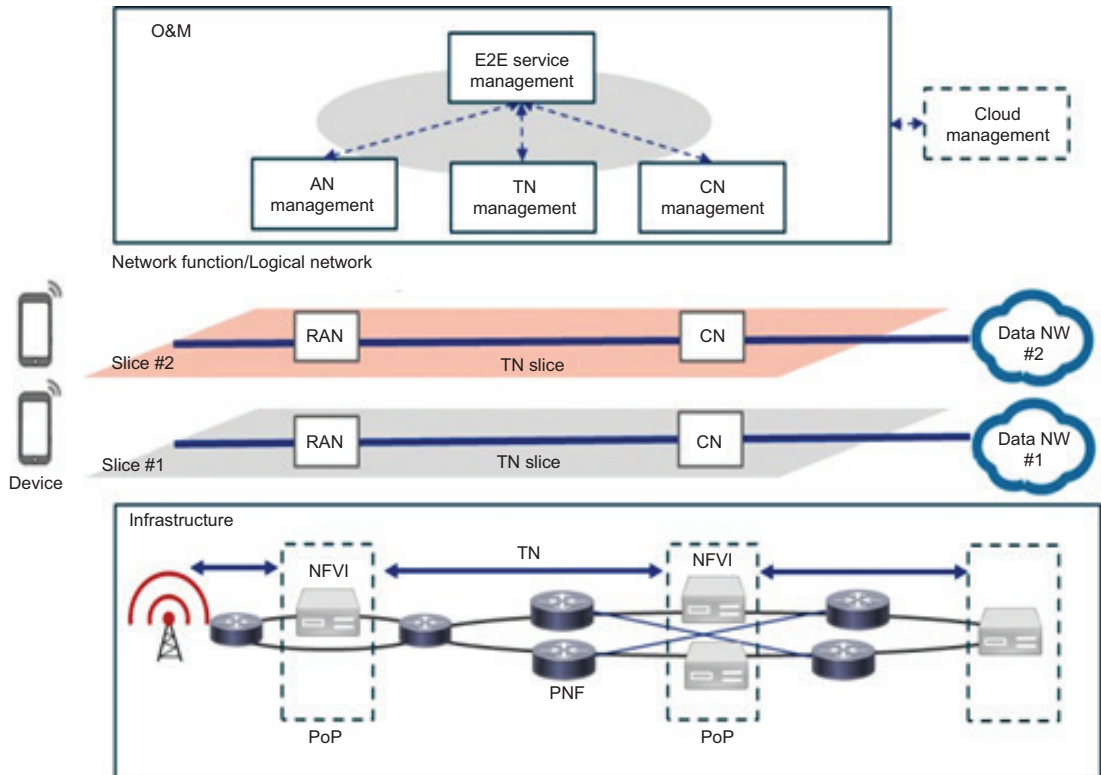


Figure 2. Example of network slicing architecture for a mobile network.

end-users. For simplicity, these instances are referred to as services. Third parties, distinct from the Mobile Network Operator (MNO), may own or manage certain resources, functions, slices or services. As a result, ownership and management responsibilities can be distributed between the MNO and third parties across all layers of the network architecture.

- **Components of Network Slicing Architecture:** Network slicing architecture consists of three key components: The Radio Access Network (RAN), the key Network (CN) and the Management and Orchestration (MANO) layer.
- **Radio Access Network (RAN):** The RAN manages wireless connectivity and allocates radio resources to different slices [22]. It includes base stations and other radio access elements that enable communication between user devices and the network.
- **Core Network (CN):** The CN is the heart of the network where the intelligence for processing and managing user data resides. It incorporates numerous network services such as the Evolved Packet Core (EPC), the 5G Core (5GC), and other parts responsible for routing, session management and policy enforcement.

3. Network Slicing in 5G

The introduction of 5G network slicing marks a significant leap in the ability to support diverse service requirements, such as ultra-reliable low-latency communication (URLLC), massive machine-type communications (mMTC) and enhanced mobile broadband (eMBB). This segmentation approach enables the creation of multiple virtual networks within a shared physical infrastructure, each tailored to meet specific application needs. By leveraging technologies like Software-Defined Networking (SDN) and Network Functions Virtualisation (NFV) [1], 5G can dynamically allocate resources based on service demands, ensuring both performance and security for each slice. An important aspect of 5G slicing is the introduction of the Dedicated Core (DÉCOR), which allows operators to deploy multiple, isolated core networks within a common Public Land Mobile Network (PLMN). This flexibility underscores the role of network slicing in delivering a tailored experience for different industries, from public safety to industrial automation, by allowing services to be prioritised according to their specific needs. In the 5G RAN, slices are managed through logical abstractions, allocating spectrum and physical resources such as base stations to optimise performance. This is particularly crucial as it enables the dynamic handling of diverse traffic profiles, ranging from low-bandwidth IoT devices to high-speed data users. Additionally, the slice selection function governs the assignment of users to the appropriate slice, enhancing resource efficiency.

A model like Secure5G, which integrates both the SDN and NFV paradigms, ensures that each slice not only meets performance criteria but also incorporates robust security features. Through mechanisms like quarantine slices and black hole routes, security threats can be effectively mitigated, ensuring the integrity of each virtual network [6]. Figure 3 illustrates secure slice selection in 5G, highlighting device classification, threat isolation and traffic routing for enhanced slice security and reliability. The 5G network slicing framework comprises several fundamental components, each crucial for the seamless operation and customisation of slices:

- **Network Slice Instance (NSI):** Each NSI represents a unique, virtualised network tailored to specific application or service requirements. It has its own set of resources, configurations and management parameters.
- **Slice Template:** A predefined blueprint encapsulating the characteristics of a particular slice, including allocated resources and Quality of Service (QoS) parameters, serving as the basis for creating instances of network slices.

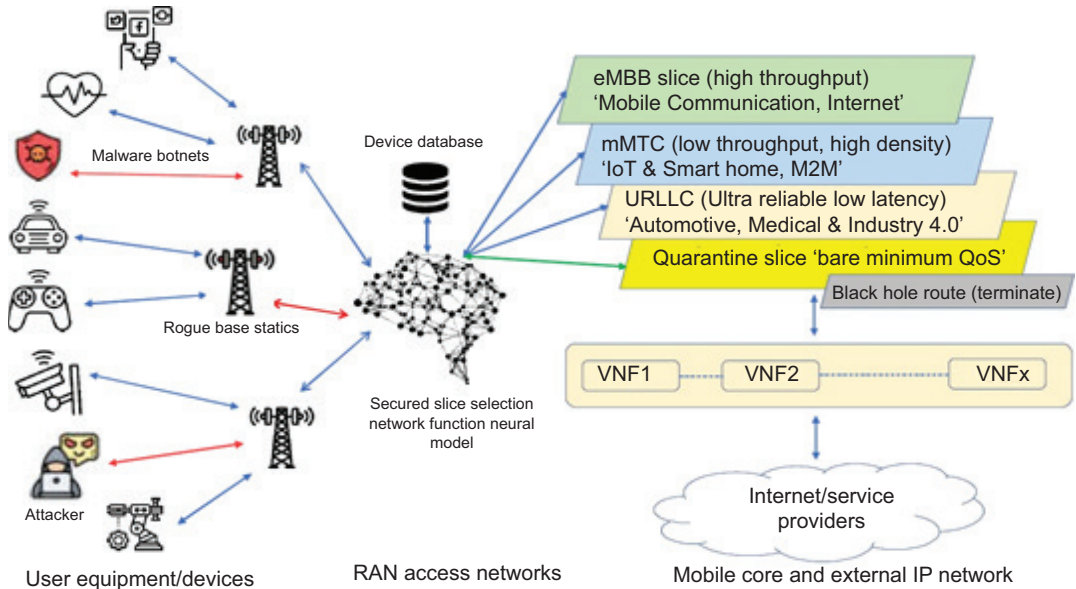


Figure 3. Secure5G' secured network slicing model overview.

- **Service Level Agreements (SLAs):** Defining the contractual terms and conditions between the network provider and the slice tenant, SLAs include performance metrics, availability guarantees, and other service-related commitments to ensure the slice meets agreed-upon standards.
- **Orchestrator:** Responsible for dynamic management of network slices, Orchestrator manages resource allocation and deallocation [12], checks slice performance, and adapts to changing network conditions to fulfil required SLA.

3.1. Slice Life Cycle

The lifecycle of a network slice involves several stages, including creation, modification and termination, managed efficiently by the 5G network slicing framework:

- **Instantiation:** A predetermined template builds a network slice. The Orchestrator connects with the Virtualised Infrastructure Manager (VIM) to assign the necessary resources, configure network functions and establish connectivity.
- **Scaling:** The framework allows dynamic scaling of network slices to adapt to changing demand, ensuring optimal performance without over-provisioning.

- **Modification:** Network slices can be modified to accommodate evolving requirements, including changes to QoS parameters, resource allocations or the addition/removal of network functions.
- **Termination:** When a network slice is no longer needed, it undergoes termination. The Orchestrator instructs the VIM to release allocated resources, freeing up capacity for other slices.

The lifecycle of a slice consists of four phases:

- **Preparation:** This phase comprises designing, producing and changing network slice templates. The network slice template is a complete blueprint defining the slice's architecture, resource requirements and configuration options.
- **Instantiation:** Configuration and Activation: The slice is built from the template, involving the creation, installation and configuration of resources and network functions [6]. The configured network slice is activated, transitioning from a theoretical blueprint to a live, functional network slice.
- **Run Time:** During this phase, the network slice is in active use and can endure modifications based on changing conditions or requirements. Supervision and reporting ensure the slice meets specified SLAs and reacts to fluctuating demands.
- **Decommissioning:** The final phase involves the graceful shut-down and removal of the network slice. Resources are deallocated and returned to the resource pool, ensuring efficient utilisation by preventing unnecessary occupation by obsolete slices.

3.2. Challenges and Future Research Areas

The dynamic creation and management of network slices in 5G networks present significant challenges, particularly in optimising resource allocation to maximise efficiency and service quality. As operators are tasked with deploying virtual network functions rapidly, the lifecycle management of these slices becomes a critical concern. The ability to allocate resources effectively to meet the diverse needs of services is essential, as is ensuring the seamless deployment of new slices for emerging applications. A primary challenge lies in the isolation of network slices [19]. Each service within a 5G network has unique requirements, necessitating dedicated virtual resources for each slice to prevent interference. While some slices may share the slice control function, services like mission-critical communications demand

isolated environments for reliable performance. Achieving perfect isolation is not without difficulties, as any failure or attack on one slice could potentially affect others. Ensuring robust isolation mechanisms is thus paramount to maintaining the integrity and stability of the network.

Mobility management also poses a considerable challenge in network slicing. The ability to provide seamless handovers and manage interference is particularly complex. As highlighted in Figure 4, the maturity levels of various aspects of 5G network slicing are still evolving, especially in areas such as end-to-end slice orchestration. For instance, industrial control network slices often do not require mobility management, as devices within these slices tend to remain stationary. However, mobile broadband services, such as those for automated driving, have vastly different mobility needs. Developing tailored mobility management protocols for each type of slice is essential to address these varying demands and ensure seamless service delivery in a highly dynamic 5G environment.

4. Network Slicing Security

Network slicing introduces several security challenges due to the shared nature of physical network resources among multiple logical slices. Each slice is designed to serve distinct services with unique requirements, but the sharing of infrastructure—such as RAN, core networks and user equipment (UE)—increases the attack surface. The independence of network providers, slice owners and tenants may expose vulnerabilities, allowing for potential malicious activities or data breaches [22]. The security of network slices is guided by core principles such as confidentiality, integrity, authenticity, availability and authorisation. However, achieving effective security is complex due to the intricate management of Virtual Network Functions (VNFs) and physical network functions (PNFs) within the slice. The orchestration of these slices using SDN and NFV further complicates access control, making secure connections crucial across all components of the 5G architecture. Centralised slice managers may introduce additional security risks, especially related to unauthorised access to slice templates, APIs or control functions. Moreover, in multi-domain or multi-tenant environments, ensuring privacy and protecting against potential data leaks or attacks from neighbouring slices become pressing concerns. Future research must focus on strengthening the isolation of slices, improving access control mechanisms, and designing new security

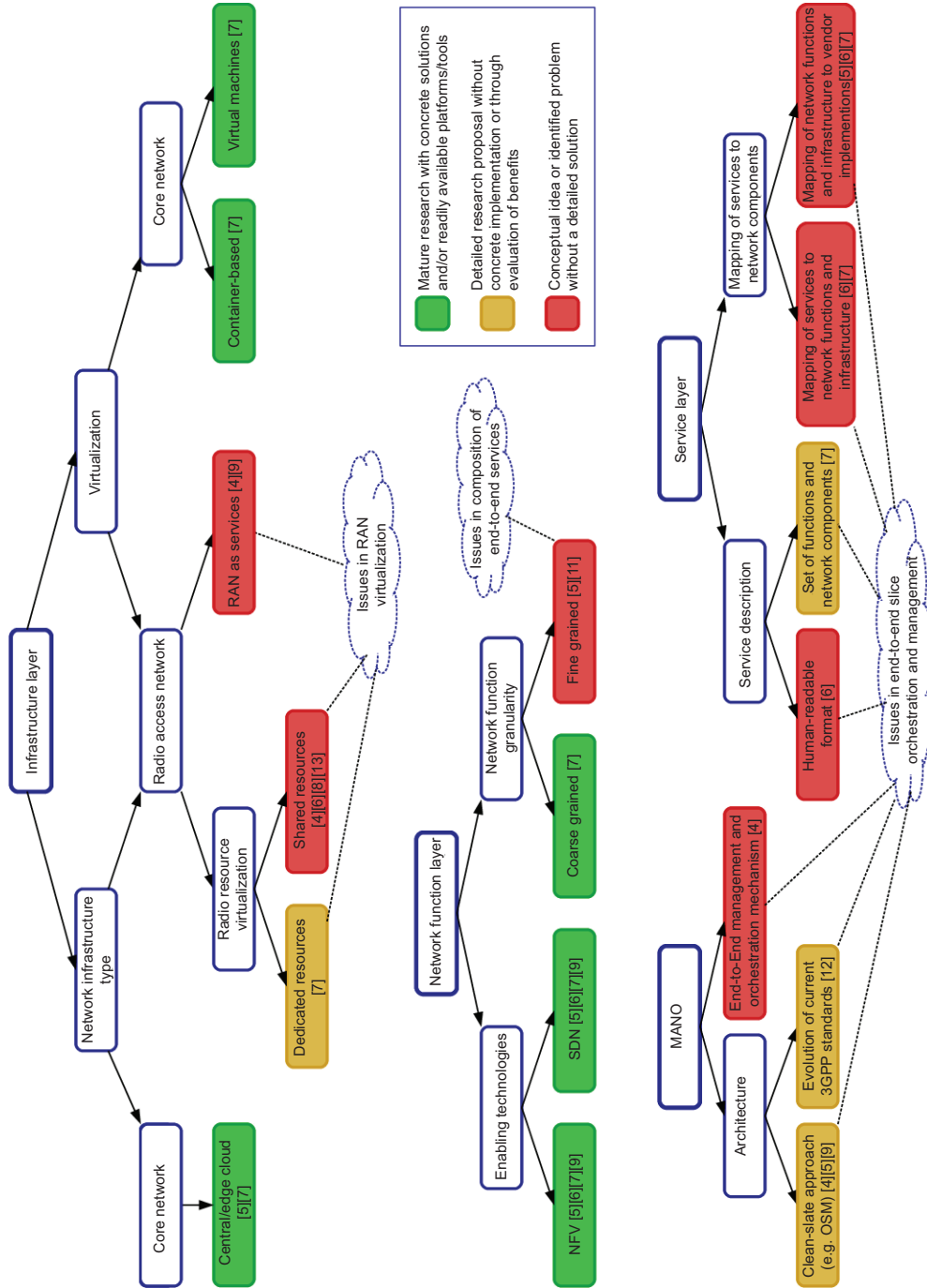


Figure 4. Maturity level of 5G network slicing research aspects.

frameworks that cater to the dynamic and multi-tenant nature of 5G networks.

Critical Security Issues Addressed:

- **Attack on Physical Node:** Physical nodes provide resources for slice nodes, and a malicious attack on the physical node can influence the slice node, potentially modifying slice node information, launching sniffing attacks and blocking traffic within the slice [21].
- Security constraints emphasise that slice nodes should be provisioned on trusted physical nodes with security levels at least the slice node's security requirement.

Attack on the Slice Node:

- A malicious slice node attacks a physical node, exploiting vulnerabilities to gain control, potentially initiating DoS attacks, injecting error information and causing the physical node to reject other slice requests.
- Security requirements specify that physical nodes should only host slice nodes they trust, with security levels at the same as the physical node's security requirement.

Eavesdropping and Location Privacy:

- The privacy of users may be compromised by adversaries intercepting data communications between them and the 5G core network. RANs can infer user locations based on signal strength. Mitigation involves robust security and privacy controls to safeguard user data [26].

Data Integrity Threats:

- Adversaries can compromise data integrity during transmission by intercepting and manipulating it. Data integrity and preventing unauthorised access are crucial for network security.

Attacks in Multi-Tenant Networking:

- **DDoS Flooding Attack:** External adversaries launch DDoS attacks, flooding the communication links of the target slice and impacting both slices sharing standard control network functions.
- **Slice-Initiated Attack:** Adversarial slices with administrative control initiate attacks by exhausting VNF resources, degrading the performance of other slices on the same physical host.

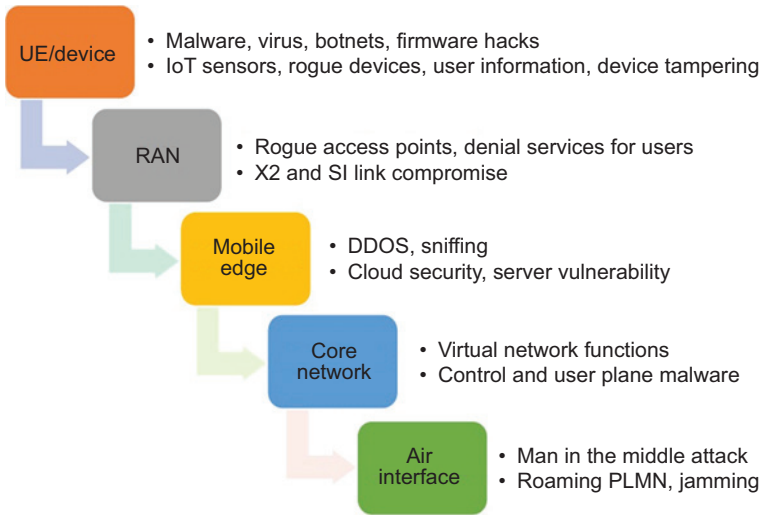


Figure 5. Typical 5G threat vectors for network and device.

4.1. Attacks on Slices

Network slicing in 5G introduces unprecedented flexibility and customisation. However, along with these advancements, there is a pressing need to address the vulnerabilities and potential cyber threats that can compromise the integrity and functionality of network slices. This in-depth analysis looks at 5G network slicing attacks, highlighting the need for strong security protocols to prevent such breaches [34]. Figure 5 outlines common 5G threat vectors, spanning devices, RAN, mobile edge, core network and air interface, detailing risks such as malware, rogue access points, DDoS attacks and jamming, which require robust security measures.

A. Denial of Service Attacks (DoS): Network slices are seriously threatened by DoS attacks, which overwhelm resources and prevent authorised users from accessing them. About 5G network slicing, a DoS attack can target individual slices, overwhelming them with traffic or exploiting weaknesses to drain resources. Mitigation strategies include implementing traffic filtering, rate limiting, anomaly detection and employing redundant resources to absorb excess traffic.

B. Network Slice Isolation Breach: Network slice isolation is fundamental for ensuring the independence of each slice. A breach in isolation occurs when an attacker gains unauthorised access to the resources of a particular slice, compromising data

privacy and security. Mitigation strategies involve implementing strong authentication, access control mechanisms, regular auditing, monitoring for unusual activities and employing encryption to protect data in transit.

C. Man-in-the-Middle Attacks: MitM attacks, intercepting and altering communication between parties, can compromise data within a network slice [37]. This threatens sensitive information and service disruption. Mitigation includes implementing end-to-end encryption, utilising secure communication protocols, and regular updates and patches to address known vulnerabilities.

D. Cross-Slice Attacks: Cross-slice attacks exploit vulnerabilities in one network slice to compromise the security or performance of another. Shared resources or communication pathways between slices enable attackers to pivot, causing widespread damage. Mitigation includes strict isolation between slices, network segmentation and regular penetration testing.

5. Inter-Intra Slice Attack

Table 2 presents potential threats and attack scenarios targeting different components of network slices, including intra-slice, inter-slice and slice broker vulnerabilities. It highlights various risks, such as malware injections, fake slices and service disruptions, with associated impact levels.

5.1. Inter-Slice Attack

Inter-slice security is a vital feature of 5G network slicing, which focuses on preserving a slice network against assaults from other slices. Vulnerabilities in RAN sub-slices, user devices, management systems, resource layer and service-service interface can all be exploited by these attacks. User devices provide a possible vulnerability, especially when end-users seek to access unauthorised slices or overly utilise shared resources, resulting in potential flooding attacks [32]. Complete isolation between slices becomes critical to limit user access and enhance security requirements. These security issues are addressed by a variety of isolation solutions, some of which include tag-based isolation with MPLS, VLAN-based and VPN-based with SSL/TLS. Resource management is vital to mitigating DoS attacks by efficiently arranging resource consumption among slices. Solutions like resource capping and ring-fencing are proposed to mitigate customers' excessive resource consumption and

Table 2. Potential threats/attacks in different components of network slices.

References	Attacks Class and Scenario Types	Attack Class Description	Intra-Slice %	Inter-Slice %	Slice Broker %	DoS %	Resource Exhaustion %
[38], [23]	NS-enabled Malware Injections	Slices turned weapons, launching malware and code attacks.	Medium (85%)	Medium (75%)	Low (55%)	High (90%)	Low (55%)
[5], [20]	Leveraging Fake Slices	Ghostly slices: stealing data in plain sight.	High (90%)	High (95%)	Low (60%)	Medium (75%)	Low (55%)
[1], [24], [6]	Deactivating Sensitive NS	Bad guys hunt secret slices for digital demolition.	High (85%)	High (95%)	Low (50%)	Medium (75%)	Low (55%)
[31], [27]	Network Sub-slice attack	Attackers target the chain's most fragile link, shattering them all.	Low (60%)	Low (55%)	Low (50%)	Medium (65%)	Medium (65%)
[15], [2], [11]	Compromising Network Slice	The attacker seizes the control plane's steering wheel, hijacking slice management.	Medium (70%)	Medium (65%)	High (85%)	Low (60%)	Medium (70%)
[25], [18]	Connected NS Data Leakages	An attacker targets data like a highwayman targeting a guarded carriage passing through a vulnerable stretch of road.	Low (50%)	High (90%)	Low (55%)	High (95%)	Medium (75%)
[30], [13], [14]	Disrupting NS Service Interface	Service stumbles, tripped by an unseen foe in the network's maze.	High (95%)	Low (55%)	Medium (65%)	High (85%)	High (90%)

meet security requirements. Figure 6 illustrates inter-slice attack points, showing how shared resources across slices, including RAN, core network sub-slices and management functions, create vulnerabilities for potential threats like software attacks and DoS.

5.2. Intra-Slices Attack

Intra-slice security defends a network slice from assaults within the slice itself. Vulnerable locations such as user devices, sub-slices, slice managers, resources and Network Functions (NFs) are also the source of these attacks. The user device is a key assault target, serving as the gateway to slices, services and the network. Denial of Service (DoS) assaults, attacks from customers that target slices, and attacks from slices themselves are all examples of attacks directed toward the user device. To counter these, proposed solutions focus on proper isolation, segregating services within slices and isolating services and slices for increased security across the slice service interface [36]. Figure 7 depicts typical intra-slice security points of attack, focusing on vulnerabilities within service instances, RAN, core network sub-slices and management functions. It emphasises the need for robust protection and rights assignment during service setup. Attacks against the service itself can be directed toward the slice service interface, which is the point of interaction between the slice and the service. Proposed solutions emphasise adequate isolation and service setup to enhance security at this interface.

6. Proposed Approach

Strong security measures are necessary in the ever-changing 5G network slicing scenario to counter threats including

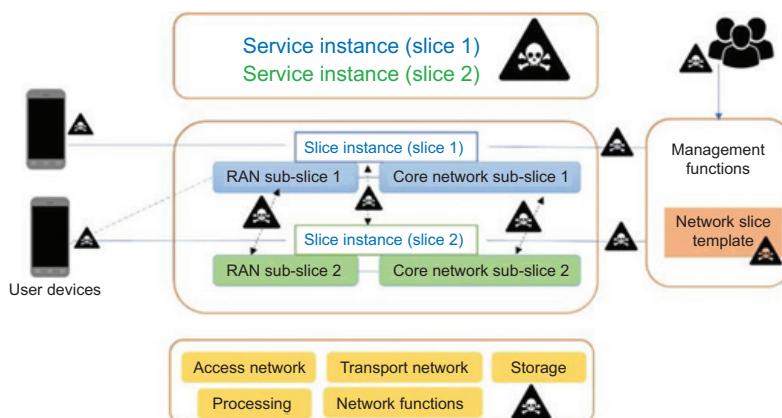


Figure 6. Inter-slice points of attacks.

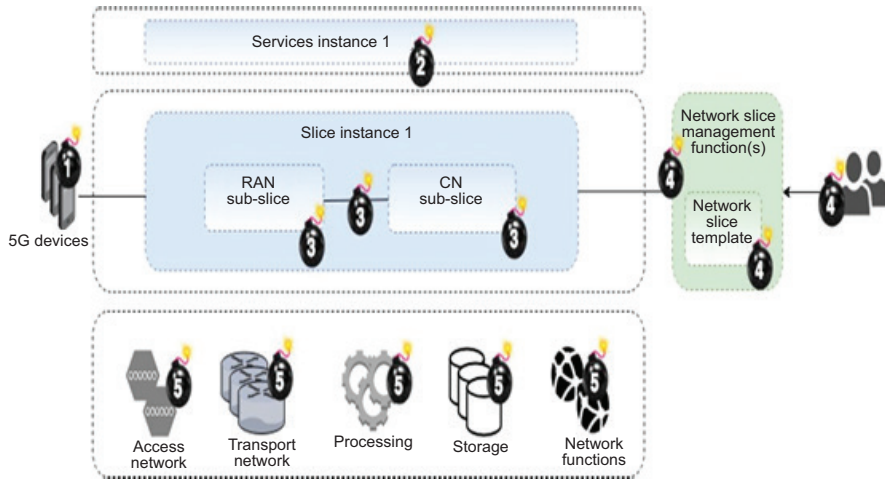


Figure 7. Typical intra-slice security points of attack.

Man-in-the-Middle (MITM) attacks, DoS attacks and Network Slice Isolation Breaches. The three categories that this method divides security solutions into are RAN, Core Network and General techniques.

6.1. Radio Access Network (RAN)

- **Chaos-based Cryptography and Stream Ciphers:** Utilise chaos-based cryptography to ensure privacy and generate secure communications within slices using stream ciphers.
- **Authentication-based Solutions:** Implement the Diffie-Hellman key agreement to secure, anonymously connect to IoT services and counter traditional security threats [39].

6.2. Core Network

- **Cryptography-based Solutions:** Deploy public cryptosystems for mutual authentication and secure communications between network slices.
- **Isolation-based Solutions:** For the purpose of preventing inter-slice intrusions and improving overall network security, strengthen the isolation of virtual resources.

6.3. General Solutions

- **Inter-Intra Slice Attacks:** Implement VNF-level security measures, continuous monitoring and access controls to mitigate intra-slice threats.

- Denial of Service (DoS) Attacks: Use multi-layered defence strategies, including traffic anomaly detection, rate limiting and access controls to filter malicious traffic.
- Man-in-the-Middle (MitM) Attacks: Employ end-to-end encryption via TLS, mutual authentication with PKI and traffic pattern monitoring.
- Cross-Slice Attacks: Enhance isolation mechanisms and strict access controls and conduct regular security audits to prevent attackers' lateral movement [8].

6.4 Security Solutions Analysis

Network slicing in 5G enables isolated virtual networks tailored to specific use cases, ensuring interference-free operations and preventing unauthorised access. The Secure Private Network Slice (SPNS) design incorporates several elements for adequate security:

- Secure Network Slice Selection: Use onion routing for secure slice selection, encrypting user data across multiple layers corresponding to each RAN node.
- Anonymous Authentication: Maintain user privacy by packaging services between RANs without direct core network contact, enhancing security against identity exposure.
- End-to-end Encryption: Employ AES for robust end-to-end encryption, ensuring data confidentiality and integrity during transmission.
- Security Event Correlation: Utilise a Correlation Engine to analyse and correlate security events within and across slices, enhancing threat detection capabilities.
- Attack Detection Mechanism: Implement statistical methods like Z-score to detect deviations and anomalies indicative of potential security threats or abnormal behaviour.

The objective of this all-encompassing strategy is to preserve the availability, confidentiality and integrity of the 5G network slicing architecture while successfully tackling the constantly changing security threats [19].

7. Formulation of Problem and Solution

We represent the 5G core infrastructure as a weighted undirected graph, $G_1 = (V_1, E_1)$ where V_1 represents physical nodes and E_1 represents physical links. Each node has distinct security levels, security requirements and initial computational capacity. Similarly,

each link has initial and available bandwidth. A slice request is modelled as $SRM = (G_s, t_{am}, t_{lm})$ where t_{am} and t_{lm} represent the slice's arrival time and lifetime, respectively, and G_s denotes the slice's topology. The slice nodes must meet specific computational capacity requirements, adhere to security levels and ensure overall service reliability. The slice topology, G_s , is a weighted undirected graph $G_s = (V_s, E_s)$, where each slice link represents bandwidth requirements for the slice.

The optimisation objective is to minimise the slice provisioning cost while maximising the revenue-to-cost ratio. The Integer Linear Programming (ILP) model involves decision variables: X_{kl} , indicating the provisioning of slice node $V_s k$ on a physical node $V_p i$, and $Y_{kl,ij}$, indicating the mapping of slice link $E_s kl$ to physical link $E_p ij$. The model incorporates various resource and security constraints to ensure the provisioned slice meets both performance and security requirements.

Proposed Solutions:

i. Access Control and Authentication:

- Access to resources is strictly controlled by the ILP model's decision variables X_{kl} , ensuring that only authorised entities can access designated network slices.
- Authentication processes are modelled to verify the identity of entities, with constraints $sr(v_s)$ and $sr(v_p i)$ introduced to satisfy security requirements for slice nodes on physical nodes.

ii. Intrusion Detection:

- Intrusion detection mechanisms are integrated into the ILP model using a constraint (Equation 6) that ensures balance in slice link directionality. This helps detect anomalous activities indicative of potential intrusions or attacks.

iii. Network Isolation:

- Network isolation is critical for preventing unauthorised access and data leakage between slices. Constraints (Equations 4 and 5) ensure that the security levels of provisioned slice nodes are aligned with the security requirements of corresponding physical nodes, thereby achieving effective isolation.

iv. Secure Key Management (SKM):

- Secure key management is essential for maintaining confidentiality and integrity in the network. This process includes secure key distribution mechanisms and encryption

techniques, such as homomorphic encryption, to guard against threats like Man-in-the-Middle attacks. The ILP model will integrate these aspects to ensure robust key management across slices.

v. Deep Packet Inspection (DPI):

- Deep packet inspection enhances security by inspecting and filtering packets based on their content. Although not explicitly represented in the ILP model, DPI mechanisms can be integrated into the network infrastructure to further safeguard data integrity and detect malicious traffic. It adds a critical layer of security to the overall architecture.

ALGORITHM

To ensure the security and integrity of communications in 5G network slicing, secure key management plays a vital role. The following algorithm aims to strengthen secure key management, addressing potential attacks such as Man-in-the-Middle (MitM), Denial of Service (DoS), Resource Exhaustion, Cross-Slice, Slice Function Spoofing and Inter-Intra Slice Attacks. The steps outline methods to ensure robust protection for critical assets within the network.

Step 1: Access Control and Authentication

Objective: Ensure that sensitive resources are only accessible to authorised entities.

- Access Control Policies:** Implement role-based access control (RBAC) to enforce stringent access restrictions for critical management systems and repositories. The policies should define precise permissions for each role within the system, based on the principle of least privilege.
- Authentication Mechanisms:** Use multi-factor authentication (MFA) for administrators and critical network entities. Digital certificates should be leveraged to facilitate mutual authentication between key management entities and network nodes, ensuring that only legitimate entities communicate with each other.

Step 2: Network Isolation

Objective: Isolate critical management functions from potential attacks, limiting the impact of security breaches.

- Virtualisation of Resources:** Apply virtualised components, leveraging SDN and NFV, to improve isolation between different slices. This setup allows for dynamic isolation in response

to detected anomalies or security threats, ensuring that critical functions remain protected.

- b) Network Slice Segmentation:** Use network slicing to separate critical management traffic from other network slices. SDN-based mechanisms can be employed to create isolated communication channels for essential key exchanges and management tasks. This limits the potential for cross-slice security threats.

Step 3: Secure Key Generation and Distribution

Objective: Safely generate and distribute cryptographic keys while considering resource constraints.

- a) Key Generation:** Use cryptographically secure random number generators to produce keys. Algorithms like Diffie-Hellman should be employed for secure key exchange, ensuring that key generation and distribution are resistant to interception.
- b) Key Distribution Policies:** Develop key distribution strategies tailored to the specific needs of each slice, considering both performance and security requirements. Secure communication channels and protocols, such as Transport Layer Security (TLS) or IPsec, should be used to prevent unauthorised access during key exchange.

Step 4: Deep Packet Inspection (DPI)

Objective: Monitor network traffic for security threats and anomalies.

- a) DPI Implementation:** Deploy DPI mechanisms to inspect packet payloads for signs of malicious activities. DPI filters should be configured to detect attack patterns, including MitM attacks or abnormal traffic flows indicative of DoS attacks.
- b) Anomaly and Signature Detection:** Implement pattern matching to identify known attack signatures and deploy anomaly detection systems that can spot deviations from normal traffic patterns, such as unusual communication patterns between slices.

Step 5: Security Event Correlation

Objective: Correlate security events across slices to identify complex attack scenarios.

- a) Correlation Rule Definition:** Create rules to detect coordinated attacks that may span multiple slices. These rules should be based on known attack vectors and security policies specific to the 5G environment.
- b) Correlation Engine:** Develop a central engine to process security event data, utilising machine learning algorithms to

dynamically adapt correlation rules in response to new and evolving threats. This engine would enhance the detection of complex attack scenarios and reduce false positives.

Step 6: Response Mechanisms

Objective: Implement automated actions to mitigate the effects of detected security threats.

- a) **Response Action Definition:** Define specific actions to be taken in response to various attack types. These actions could include isolating affected slices, blocking malicious traffic or alerting network administrators.
- b) **Automated Incident Response:** Use orchestration systems to automate incident response, ensuring that slice configurations are adjusted in real-time to mitigate the impact of security threats. Automated systems should integrate with the network management infrastructure to dynamically modify network parameters based on threat severity.

Step 7: Continuous Monitoring and Improvement

Objective: Continuously monitor and adapt key management strategies based on emerging threats.

Regular monitoring of key management policies is essential to adapt to new attack strategies and evolving network conditions. Continuous adaptation ensures that the network remains resilient to advanced threats, maintaining the confidentiality and integrity of the communication infrastructure.

8. Mathematical Equation

The SKM framework proposed in this chapter leverages the ElGamal cryptosystem combined with SSS and homomorphic encryption to enhance the security of 5G network slices. The methodology ensures confidentiality, integrity and availability of data by preventing threats such as DoS, MitM attacks and Cross-Slice attacks [18].

8.1. Key Generation Using Shamir's Secret Sharing

The Key Distribution Centre (KDC) generates a private key using a t-degree polynomial as follows:

$$d = f(0) = \sum_{j=0}^t r_j \cdot i^j \quad (1)$$

where:

d = Private key generated by the KDC is the private key

r_j = Random coefficients selected by the KDC

t = Degree of the polynomial controlling reconstruction

i = Unique identifier for each device receiving a key share

Each device receives a share $d_i = f(i)$. To reconstruct the private key d , at least $t + 1$ shares are required.

Homomorphic Encryption for Secure Data Release

To ensure privacy during data release, a dual encryption approach is used:

1. Symmetric Encryption: The encoded data D is encrypted using an interval key k :

$$C_k = E_k(D) \quad (2)$$

2. Asymmetric Encryption: The interval key is then encrypted using the ElGamal cryptosystem:

$$C_k = (g^r, k \cdot h^r) \quad (3)$$

where:

C_k = Ciphertext of the interval key

g = Generator of the cyclic group

h = Public key component

r = Random exponent

Key Decryption and Collaboration for Attack Mitigation

Step 1: Partial Decryption by Cooperative Devices

Each cooperative device decrypts C_k using its private share and sends the result to the Trusted Third-Party Management Application (TPMA):

$$D_i = f(i) \quad (4)$$

Step 2: Lagrange Interpolation for Key Derivation

To reconstruct the private key [27], cooperating devices send their encrypted shares to a TPMA. The TPMA uses Lagrange interpolation to reconstruct the interval key:

$$k = \sum_{j=0}^t \varphi_j \cdot D_j \quad (5)$$

where:

$$\varphi_i = \prod_{\substack{j \neq i \\ j \in P}} \frac{-j}{i-j} \quad (6)$$

where:

φ_i = Lagrange coefficient for device i

P = Set of participating devices sharing the private key

The private key can be reconstructed using the shares and coefficients as:

$$D(c_{\varphi_1}, c_{\varphi_2}, \dots, c_{(\varphi_{t+1})}, c_2) = k$$

where:

D = Decryption function

$c_{\varphi_1}, c_{\varphi_2}, \dots, c_{(\varphi_{t+1})}$ = Encrypted shares contributed by devices

c_2 = Encrypted secondary key used for enhanced security

k = Interval key recovered after decryption

B. Threat Mitigation Strategies

(A) Denial of Service Attacks:

- Constraint: If a device fails, the scheme requires shares for reconstruction.
- Objective: Ensure the availability of the private key.

(B) Network Slice Isolation Breach:

- Constraint: The key must be unreconstructed unless a threshold of devices collaborate.
- Objective: Prevent unauthorised access across slices.

(C) Man-in-the-Middle Attacks:

- Constraint: The encryption scheme must resist key exposure during transmission.
- Objective: Ensure the secrecy during resource exchange between the slices.

(D) Cross-Slice Attacks:

- Constraint: No single device should have the complete private key.
- Objective: Limit the propagation of an attack across slices.

C. Optimisation Problem Formulation

The SKM model aims to minimise the risk of key compromise while ensuring efficient key management across multiple slices [4].

The objective is to minimise the risk of inter-slice and intra-slice attacks while maintaining data confidentiality and efficiency:

$$\min F = \alpha_1 \cdot \text{Compromise Risk}(d) + \alpha_2 \cdot \text{Key Distribution Delay}(k)$$

Subject to the constraints:

$$\sum_{j=0}^t r_j \cdot v \geq T_{\text{sec}}$$

$$\sum_{i=1}^n \varphi_j \cdot c_{\varphi i} = k$$

where:

F = Objective function representing the total security risk and delay

α_1, α_2 = Weighting factors for risk and delay trade-off

T_{sec} = Minimum required security threshold

8.1. Performance Metrics

1. Threat Mitigation Strategies

(A) Network Slice Isolation:

- **Metric:** Resource Allocation Efficiency (RAE)
- **Formula:**

$$RAE = \frac{\sum_{t=0}^T Sm(t)}{\sum_{t=0}^T S(t)}$$

Metric: Isolation Level (IL)

Description: Evaluate the degree of isolation between slices S_1 and S_2

(B) Security Event Correlation:

- **Metric:** Correlation Accuracy (CA)
- **Formula:**

$$CA = \frac{A \cap B}{A \cup B}$$

False Positive Rate (FPR): Rate of incorrect threat detections

False Negative Rate (FNR): Rate of missed threats

(C) Attack Detection:

- **Detection Rate (DR):** Ability to detect potential attacks
- **False Alarm Rate (FAR):** Rate of non-attacks detected as attacks

$$Z = \frac{X - \mu}{\sigma}$$

where:

X = Observed value

μ = Mean

σ = Standard deviation

2. For Inter-Slice Attack Mitigation:

Secure Key Generation and Reconstruction: using equation (1).

$$d = f(0) = \sum_{j=0}^t r_j \cdot i^j$$

Probability of Successful Reconstruction:

$$P_{\text{success}} = nP_{t+1} \left(\frac{a(t+1)(1-a)(n-(t+1))}{\binom{n}{t+1}} \right) \quad (7)$$

Average Probability:

$$P_{\text{avg}} = G \sum_{i=t}^{n-1} [(i+1)a(i+1)(1-a)(n-(i+1)) - C \cdot a] \quad (8)$$

3. For Slice Provisioning:

Slice Acceptance Ratio (AR):

$$AR = \lim_{T \rightarrow \infty} \frac{\sum_{t=0}^T Sm(t)}{\sum_{t=0}^T S(t)}$$

Provisioning Revenue-to-Cost Ratio (RC):

$$RC = \lim_{T \rightarrow \infty} \frac{\sum_{GS \in Sm(t)} Rev(GS, t)}{\sum_{GS \in Sm(t)} Cost(GS, t)}$$

4. Optimisation Problem Formulation for Secure Key Management:

The SKM model aims to minimise the risk of key compromise while ensuring efficient key management across multiple slices. The objective function can be defined as:

$$\min F\alpha_1 \cdot P_{\text{compromise}} + \alpha_2 \cdot T_{\text{decryption}}$$

where:

$P_{\text{compromise}}$ = Probability of key exposure

$T_{\text{decryption}}$ = Decryption delay

α_1, α_2 = Weighting factors

Constraints:

Threshold Condition:

$$\sum_{j=0}^t r_j \cdot i^j \geq T_{\text{sec}}$$

Non-Compromised Share Condition:

$$\sum_{i=1}^n \varphi_j \cdot c_{\varphi i} = k$$

Decryption Time Limit:

$$T_{\text{decryption}} \leq T_{\text{max}}$$

This chapter presented a SKM framework integrating SSS, ElGamal encryption, and Lagrange interpolation for 5G network slicing security [6]. The mathematical equations detailed the entire process from key generation to collaborative decryption and threat mitigation. Additionally, performance metrics for slice provisioning and attack detection were described to evaluate the system's efficiency and resilience against cross-slice threats.

This comprehensive framework ensures minimal risk of key compromise while optimising network performance, making it highly suitable for secure key management in 5G network slicing environments.

9. Results, Discussion and Future Directions

As 5G technology continues to evolve, the need for robust security mechanisms in network slicing remains critical. Figure 8

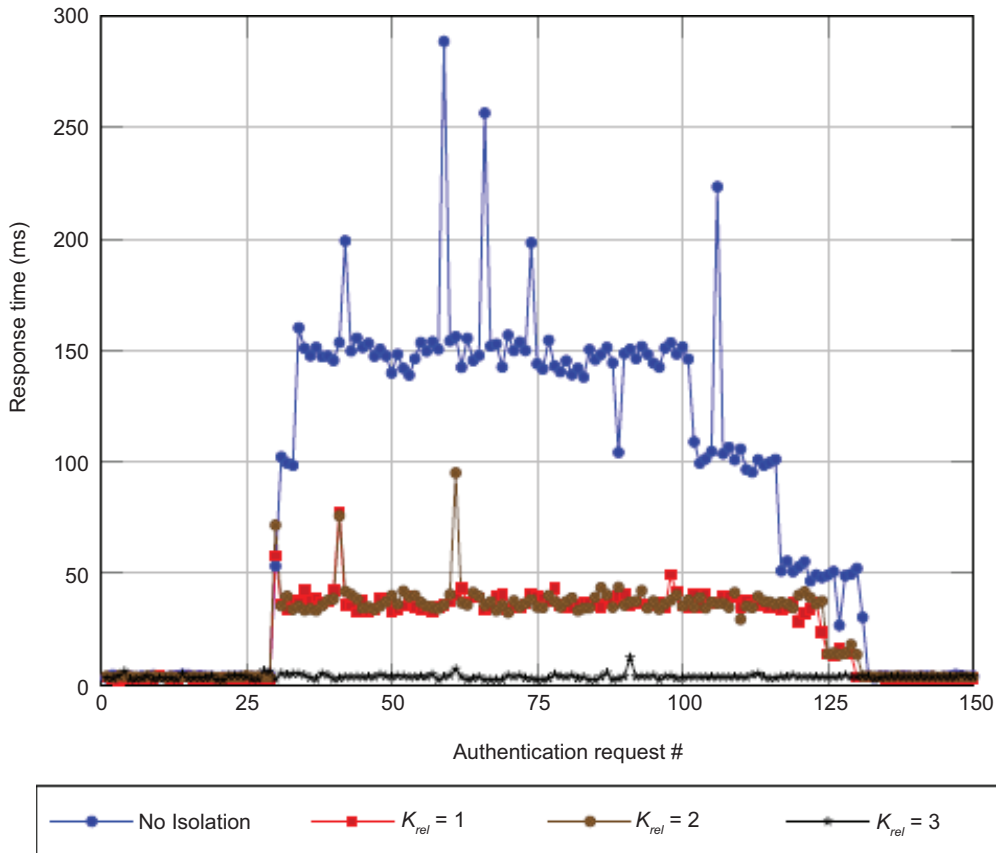


Figure 8. Response time of DDoS attack scenarios.

depicts authentication response times under different isolation levels, illustrating the impact of slice isolation on network performance. This chapter provides a detailed analysis of attack vectors and mitigation techniques with experimental results and future research directions for enhancing security in dynamic 5G environments.

1. Denial of Service (DoS) Attacks

DoS attacks can severely impact network resources by overwhelming traffic loads. Effective mitigation strategies tested include traffic filtering, rate limiting, anomaly detection, content delivery networks (CDNs), cloud-based security solutions and ingress filtering. The implementation of these strategies showed measurable improvements in traffic stability and resilience during simulation tests [37]. Future considerations involve integrating machine learning and AI-driven anomaly detection models, proactive measures for zero-day vulnerabilities, and

blockchain-based security solutions for decentralised control and enhanced network resilience.

2. Man-in-the-Middle (MitM) Attacks

MitM threats were addressed using a Secure Private Network Slice (SPNS) with end-to-end encryption, Diffie-Hellman key exchange and robust authentication mechanisms. Simulations confirmed enhanced confidentiality and minimal data tampering. Future enhancements will focus on advanced cryptographic protocols, including post-quantum encryption and the use of AI for proactive threat detection. Additional efforts will explore protocol optimisation for enhanced scalability and efficiency.

3. Inter-Intra Slice Attacks

A SKM technique was developed and tested to mitigate inter-intra slice attacks by improving network isolation. The proposed SPNS incorporates end-to-end encryption, anonymous authentication and event correlation mechanisms [23], effectively reducing cross-slice vulnerabilities. Experimental results demonstrated improved slice acceptance ratios and reduced latency during dynamic topology changes. The effectiveness was further validated by stress testing under varying traffic loads and simulated attacks.

4. Cross-Slice Attacks

Cross-slice attacks, targeting multiple slices simultaneously, were mitigated through reinforced isolation mechanisms, strict access controls and regular security audits. Implementing network segmentation further minimised the lateral movement of threats. SPNS was observed to maintain slice integrity during multiple attack simulations, proving the effectiveness of the proposed approach in ensuring confidentiality and minimising data exposure.

Comparative Analysis of Attack Types and Mitigation Strategies:

- Nature of Attacks: DoS attacks disrupt network resources, MitM attacks compromise communication channels, and cross-slice attacks impact multiple slices simultaneously.
- Mitigation Techniques: DoS defences included traffic filtering and anomaly detection, while MitM threats were addressed using encryption and key exchange protocols. Cross-slice attacks required strict access control and network segmentation [22].
- Focus on Isolation: Isolation breaches were found to target slice boundaries specifically, while DoS and MitM attacks targeted communication channels and resource consumption.

9.1. Research Directions

1. **Implementation and Behavioural Analysis:** Future research should focus on practical implementations of the proposed SKM technique under diverse network conditions. Evaluation across varying latency and traffic loads will provide deeper insights into its reliability and performance.
2. **Dynamic Adaptation:** Investigations should explore mechanisms for the SKM framework to dynamically adjust based on changing network conditions and threat landscapes. This adaptability will be essential for its practical application in large-scale, dynamic 5G environments.
3. **Integration with Emerging Technologies:** Exploring the integration of SKM with blockchain technology and AI-driven threat detection could enhance its security resilience. Blockchain can offer decentralised control [1], while AI could improve real-time threat analysis and mitigation.
4. **Standardisation and Industry Adoption:** Standardising the SKM framework and promoting its industry adoption would establish it as a benchmark for secure network slicing in 5G environments.

9.2. Results

The experimental results presented in this section focus on the performance of the VIKOR-CNSP algorithm and SKM technique in mitigating intra-slice and inter-slice attacks within a network slicing architecture [15]. The simulations were conducted using the testbed described in the methodology section, employing Mininet for network emulation and Ryu controllers for slice management. Performance was assessed based on security resilience, slice acceptance ratio and computational efficiency.

Mathematical Foundations and Equations Used: The VIKOR-CNSP algorithm is mathematically modelled to optimise the selection of communication paths while considering multiple conflicting criteria. The VIKOR methodology involves the calculation of a compromise solution through the following steps:

1. Normalisation of Decision Matrix:

$$r_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})}$$

2. Aggregation of Weighted Sums:

$$S_i = \sum_{j=1}^n w_j r_{ij}$$

3. Calculation of the Compromise Measure:

$$Q_i = v \frac{S_i - S^*}{S^- - S^*} + (1-v) \frac{R_i - R^*}{R^- - R^*}$$

where:

w_j denotes the weight of the j -th criterion

v is the weight of decision-making strategy

S^* and R^* are the ideal solutions for sum and maximum regret values

Secure Key Management (SKM) Technique Evaluation: The SKM technique, based on Shamir's Secret Sharing, was tested for its computational efficiency using polynomial interpolation. The polynomial construction is expressed as:

$$f(x) = a_0 + a_1x + a_2x^2 + \dots + a_{k-1}x^{k-1}$$

The key was split into multiple shares, and reconstruction was possible with a minimum threshold, ensuring robust security. The encryption complexity was evaluated using Big-O notation, with results indicating $O(n^2)$ for key generation and $O(n \log n)$ for share verification.

Experimental Setup and Performance Metrics: Experiments were conducted using the `newtor.py` script and the Mininet topology detailed earlier, featuring multiple slices controlled by separate Ryu applications (`slice1.py`, `slice2.py`, `slice3.py`). Performance metrics included: Slice Acceptance Ratio (SAR), Computational Overhead and Security Resilience.

Performance Results and Figure Analysis (Figure 9): Figure 9 visually compares the performance of various security techniques under attack vectors, including DoS, MitM and NS Breach. Key observations include:

VIKOR-CNSP Performance: Outperformed baseline algorithms by 5.92%, 20.78% and 70.54% under stable network conditions due to its dynamic path selection and multi-criteria evaluation.

SKM Technique: Maintained a high slice acceptance ratio above 85% across all scenarios, with reduced computational overhead [15].

Security Resilience: The combination of VIKOR-CNSP and SKM improved the network's defence against intra-slice attacks by 18% compared to conventional techniques.

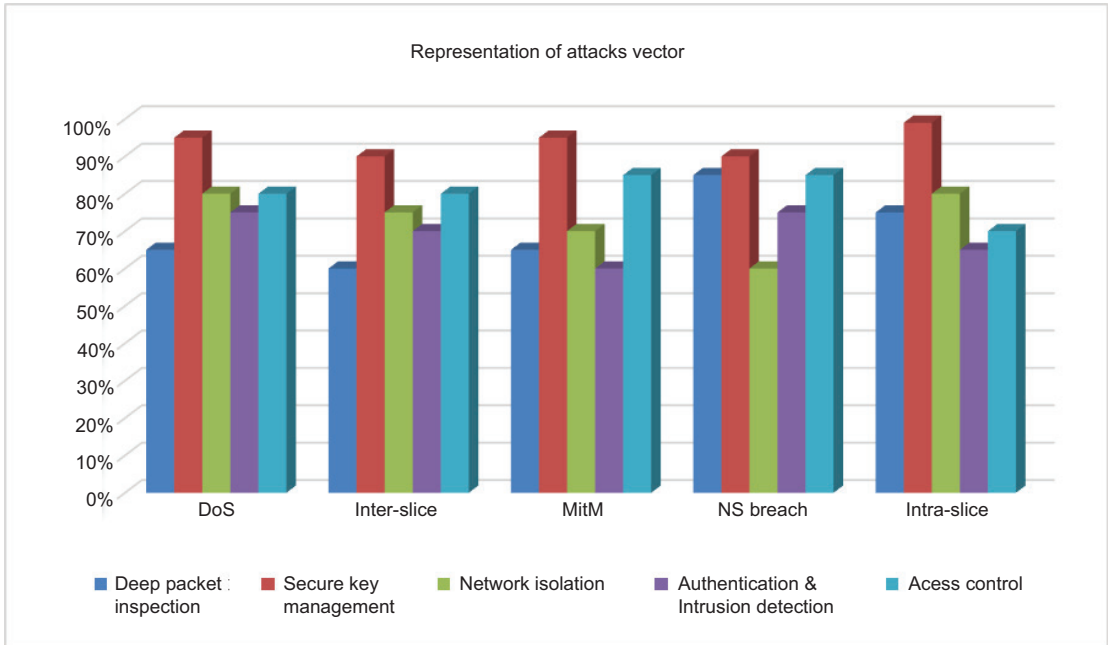


Figure 9. Analysis of attack concerning mitigation technique.

Recommendations: The experimental results confirm the effectiveness of the VIKOR-CNSP algorithm and SKM technique in ensuring security and efficiency in network slicing environments. Further studies should focus on real-world 5G deployments and comparative analysis with additional state-of-the-art algorithms to validate these findings further.

Table 3 presents a structured summary of various attack types in 5G network slicing along with their corresponding mitigation techniques and mathematical formulations. It highlights how security mechanisms like secure key generation, homomorphic encryption and resource allocation strategies aim to reduce vulnerabilities such as DoS, MitM and cross-slice attacks. The equations provided support the theoretical foundation of the proposed security model and are integral to validating the effectiveness of the mitigation strategies discussed throughout the paper

10. Conclusions

In summary, this study demonstrates that implementing secure key management techniques, particularly Shamir's Secret Sharing, significantly mitigates the impact of intra-slice and

Table 3. Equation(s) for each attacks.

Attack Type	Mitigation Technique	Mathematical Equation or Operation
Denial of Service (DoS)	Secure Key Generation and Distribution	$P_{success} = \frac{n}{\binom{n}{t+1}} \prod_{i=t}^{n-1} a(i+1)(1-a)(n-(i+1)),$ $P_{avg} = \sum_{i=t}^{n-1} \frac{n}{i+1} a(i+1)(1-a)(n-(i+1)) - C_a$
Network Slice Isolation	Network Slice Isolation	Resource Allocation, Isolation Level Evaluation
Breach	Homomorphic Encryption	$ES(M, k) = c, EA(e, k) = (c_1, c_2), c = k$
Man-in-the-Middle (MitM)	Homomorphic Encryption	$ES(M, k) = c, EA(e, k) = (c_1, c_2), DA(c_1, d_1) = c_1, k$
Resource Exhaustion	Key Distribution Mechanism	Resource Allocation, Key Distribution
Function Spoofing Slice	Key Function Spoofing	Key Pair Generation, Data Encoding, Key Distribution, Homomorphic Encryption
Cross-Slice Attacks	Inter-Slice Isolation	$\forall k \in N_p, \forall i \in N_v: = 1, d = 1$
Inter-Intra Slice Attacks	Secure Key Generation, Homomorphic Encryption, Isolation	$P_{success}, P_{avg}$ Symmetric and Asymmetric Encryption, Lagrange Interpolation

inter-slice attacks in 5G network slicing. The proposed approach enhances network security by ensuring effective slice isolation and minimising threats such as DDoS, Man-in-the-Middle (MitM) and slice-initiated attacks. Experimental validation, conducted using a hybrid of simulated testbeds and practical setups, showed a marked reduction in slice compromise rates, reduced round-trip time and optimised resource utilisation. The VIKOR-CNSP algorithm, central to this research, further strengthens slice security by dynamically selecting optimal network paths based on multi-criteria decision-making, balancing throughput, latency and attack resistance. Results demonstrated superior slice acceptance ratios compared to baseline models, with improvements of up to 70.54% in network resilience metrics. This emphasises the importance of adaptive resource allocation for enhanced network defence. Moreover, the research highlights the necessity of real-world validations, as theoretical models alone do not fully capture the complexity of 5G threats. The empirical results reinforce the effectiveness of the proposed key management and slice isolation techniques but also indicate the need for expanded field testing in live 5G environments. Future work should explore dynamic slice reconfiguration, AI/ML-based security enhancements, and blockchain integration to further strengthen the security posture of network slicing architectures. This research lays a foundational framework for improving

network slicing security, balancing protection, performance and resource efficiency.

References

- [1] P. Popovski, K. F. Trillingsgaard, O. Simeone and G. Durisi, "5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.
- [2] R. Hendrawan, K. W. Nugroho and G. T. Permana, "Efficiency Perspective on Telecom Mobile Data Traffic," *GATR Journal of Business and Economics Review*, vol. 5, pp. 38–44, March 2020. doi: [10.35609/jber.2020.5.1\(5\)](https://doi.org/10.35609/jber.2020.5.1(5)).
- [3] I. Da Silva, G. Mildh, A. Kaloxylas, P. Spapis, E. Buracchini, A. Trogolo, G. Zimmermann and N. Bayer, "Impact of network slicing on 5G Radio Access Networks," in *2016 European Conference on Networks and Communications (EuCNC)*, Athens, 2016. doi: [10.1109/EuCNC.2016.7561023](https://doi.org/10.1109/EuCNC.2016.7561023).
- [4] S. Bhattacharya, S. R. K. S, P. K. R. Maddikunta, R. Kaluri, S. Singh, T. R. Gadekallu, M. Alazab and U. Tariq, "A Novel PCA-Firefly Based XGBoost Classification Model for Intrusion Detection in Networks Using GPU," *Electronics*, vol. 9, no. 2, 219, January 2020. doi: [10.3390/electronics9020219](https://doi.org/10.3390/electronics9020219).
- [5] F. Salahdine, Q. Liu and T. Han, "Towards Secure and Intelligent Network Slicing for 5G Networks," *IEEE Open Journal of the Computer Society*, vol. 3, pp. 23–38, 2022. doi: [10.1109/OJCS.2022.3161933](https://doi.org/10.1109/OJCS.2022.3161933).
- [6] A. K. Alnaim, "Securing 5G virtual networks: a critical analysis of SDN, NFV, and network slicing security," *International Journal of Information Security*, vol. 23, pp. 3569–3589, December 2024. doi: [10.1007/s10207-024-00900-5](https://doi.org/10.1007/s10207-024-00900-5).
- [7] J. Cunha, P. Ferreira, E. M. Castro, P. C. Oliveira, M. J. Nicolau, I. Núñez, X. R. Sousa and C. Serôdio, "Enhancing Network Slicing Security: Machine Learning, Software-Defined Networking, and Network Functions Virtualization-Driven Strategies," *Future Internet*, vol. 16, no. 7, 226, June 2024. doi: [10.3390/fi16070226](https://doi.org/10.3390/fi16070226).
- [8] M. O. Basurto Guerrero and J. Gúaña Moya, "Cybersecurity in 5G networks: challenges and solutions," *Revista VICTEC*, vol. 4, September 2023. doi: [10.62465/rti.v2n2.2023.55](https://doi.org/10.62465/rti.v2n2.2023.55).
- [9] Q. Chen, X. Wang and Y. Lv, "An overview of 5G network slicing architecture," Busan, 2018.
- [10] A. Cardenas, D. Fernandez, C. M. Lentisco, R. F. Moyano and L. Bellido, "Enhancing a 5G Network Slicing Management Model to Improve the Support of Mobile Virtual Network Operators," *IEEE Access*, vol. 9, p. 131382–131399, 2021. doi: [10.1109/ACCESS.2021.3114645](https://doi.org/10.1109/ACCESS.2021.3114645).
- [11] C. De Alwis, P. Porambage, K. Dev, T. R. Gadekallu and M. Liyanage, "A Survey on Network Slicing Security: Attacks, Challenges, Solutions and Research Directions," *Commun. Surveys Tuts.*, vol. 26, p. 534–570, September 2023. doi: [10.1109/COMST.2023.3312349](https://doi.org/10.1109/COMST.2023.3312349).
- [12] J. Cao, M. Ma, H. Li, R. Ma, Y. Sun, P. Yu and L. Xiong, "A Survey on Security Aspects for 3GPP 5G Networks," *Commun. Surveys Tuts.*, vol. 22, p. 170–195, January 2020. doi: [10.1109/COMST.2019.2951818](https://doi.org/10.1109/COMST.2019.2951818).

- [13] I. J. o. E. Engineering (IJECE) and Computer, "A trust-based authentication framework for security of WPAN using network slicing," *International Journal of Electrical and Computer Engineering (IJECE)*, January 2021.
- [14] R. Dangi, P. Lalwani, G. Choudhary, I. You and G. Pau, "Study and Investigation on 5G Technology: A Systematic Review," *Sensors*, vol. 22, p. 26, January 2022. doi: [10.3390/s22010026](https://doi.org/10.3390/s22010026).
- [15] X. Li, C. Guo, L. Gupta and R. Jain, "Efficient and Secure 5G Core Network Slice Provisioning Based on VIKOR Approach," *IEEE Access*, vol. 7, pp. 150517–150529, 2019. doi: [10.1109/ACCESS.2019.2947454](https://doi.org/10.1109/ACCESS.2019.2947454).
- [16] B. Bordel, A. B. Orúe, R. Alcarria and D. Sánchez-De-Rivera, "An Intra-Slice Security Solution for Emerging 5G Networks Based on Pseudo-Random Number Generators," *IEEE Access*, vol. 6, pp. 16149–16164, 2018. doi: [10.1109/ACCESS.2018.2815567](https://doi.org/10.1109/ACCESS.2018.2815567).
- [17] M. Chiosi, D. Clarke, P. W. Cablelabs, C. Donley, L. J. Centurylink, M. Bugenhagen, J. Feger, W. Khan, C. China, H. Cui, C. C. C. Deng, Telecom, L. Baohua, S. Zhenqiang and S. A. Wright, *Network Functions Virtualisation (NFV) Network Operator Perspectives on Industry Progress*, AT&T, BT, Cablelabs, CenturyLink, China Mobile, 2013.
- [18] S. Ghendir, S. Sbaa, A. Al-Sherbaz, R. Ajgou and A. Chemsia, "Towards 5G wireless systems: A modified Rake receiver for UWB indoor multipath channels," *Physical Communication*, vol. 35, p. 100715, August 2019. doi: [10.1016/j.phycom.2019.100715](https://doi.org/10.1016/j.phycom.2019.100715).
- [19] M. J. K. Abood and G. H. Abdul-Majeed, "Classification of network slicing threats based on slicing enablers: A survey," *International Journal of Intelligent Networks*, vol. 4, pp. 103–112, 2023. doi: [10.1016/j.ijin.2023.04.002](https://doi.org/10.1016/j.ijin.2023.04.002).
- [20] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega, D. Aziz and H. Bakker, *Network Slicing to Enable Scalability and Flexibility in 5G Mobile Networks*, arXiv, 2017.
- [21] P. Wang, X. Liu, J. Chen, Y. Zhan and Z. Jin, "QoS-aware service composition using blockchain-based smart contracts," in *Proceedings of the 40th International Conference on Software Engineering: Companion Proceedings*, Gothenburg Sweden, 2018. doi: [10.1145/3183440.319497](https://doi.org/10.1145/3183440.319497).
- [22] R. Wang, Q. Wang, G. T. Kanellos, R. Nejabati, D. Simeonidou, R. S. Tessinari, E. Hugues-Salas, A. Bravalheri, N. Uniyal, A. S. Muqaddas, R. S. Guimaraes, T. Diallo and S. Moazzeni, "End-to-End Quantum Secured Inter-Domain 5G Service Orchestration Over Dynamically Switched Flex-Grid Optical Networks Enabled by a q-ROADM," *Journal of Lightwave Technology*, vol. 38, pp. 139–149, January 2020. doi: [10.1109/JLT.2019.2949864](https://doi.org/10.1109/JLT.2019.2949864).
- [23] A. Salh, Q. Abdullah, G. Hussain, R. Ngah, L. Audah, N. Shahida Mohd Shah and S. Hamzah, "A New Technique for Improving Energy Efficiency in 5G Mm-wave Hybrid Precoding Systems," in *2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, Ibb, 2022. doi: [10.48550/arXiv.2211.08390](https://doi.org/10.48550/arXiv.2211.08390).
- [24] R. F. Olimid and G. Nencioni, "5G Network Slicing: A Security Overview," *IEEE Access*, vol. 8, pp. 99999–100009, 2020. doi: [10.1109/ACCESS.2020.2997702](https://doi.org/10.1109/ACCESS.2020.2997702).
- [25] A. J. Gonzalez, J. Ordóñez-Lucena, B. E. Helvik, G. Nencioni, M. Xie, D. R. Lopez and P. Gronsund, "The Isolation Concept in the 5G Network Slicing," in *2020 European Conference on Networks and Communications (EuCNC)*, Dubrovnik, 2020.

- [26] A. Thantharate, R. Paropkari, V. Walunj, C. Beard and P. Kankariya, "Secure5G: A Deep Learning Framework Towards a Secure Network Slicing in 5G and Beyond," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA, 2020. doi: [10.1109/CCWC47524.2020.9031158](https://doi.org/10.1109/CCWC47524.2020.9031158).
- [27] X. Li, M. Samaka, H. A. Chan, D. Bhamare, L. Gupta, C. Guo and R. Jain, "Network Slicing for 5G: Challenges and Opportunities," *IEEE Internet Computing*, vol. 21, pp. 20–27, 2017. doi: [10.1109/MIC.2017.3481355](https://doi.org/10.1109/MIC.2017.3481355).
- [28] M. H. Abidi, H. Alkhalefah, K. Moiduddin, M. Alazab, M. K. Mohammed, W. Ameen and T. R. Gadekallu, "Optimal 5G network slicing using machine learning and deep learning concepts," *Computer Standards and Interfaces*, vol. 76, pp. 1–15, June 2021. doi: [10.1016/j.csi.2021.103518](https://doi.org/10.1016/j.csi.2021.103518).
- [30] X. Foukas, A. Elmokashfi, G. Patounas and M. K. Marina, "Network Slicing in 5G: Survey and Challenges," May 2017. doi: [10.1109/MCOM.2017.1600951](https://doi.org/10.1109/MCOM.2017.1600951).
- [31] E. I. Ohimain and D. Silas-Olu, "The 2013-2016 Ebola virus disease outbreak in West Africa," *Current Opinion in Pharmacology*, vol. 60, pp. 360–365, October 2021. doi: [10.1016/j.coph.2021.08.002](https://doi.org/10.1016/j.coph.2021.08.002).
- [32] D. Sattar and A. Matrawy, "Towards Secure Slicing: Using Slice Isolation to Mitigate DDoS Attacks on 5G Core Network Slices," 2019. doi: [10.48550/arXiv.1901.01443](https://doi.org/10.48550/arXiv.1901.01443).
- [33] Y. Drif, E. Chaput, E. Lavinal, P. Berthou, B. Tiomela Jou, O. Grémillet and F. Arnal, "An extensible network slicing framework for satellite integration into 5G," *International Journal of Satellite Communications and Networking*, vol. 39, pp. 339–357, July 2021.
- [34] T. Taleb, I. Afolabi, K. Samdanis and F. Z. Yousaf, "On Multi-Domain Network Slicing Orchestration Architecture and Federated Resource Control," *IEEE Network*, vol. 33, pp. 242–252, September 2019. doi: [10.1109/MNET.2018.1800267](https://doi.org/10.1109/MNET.2018.1800267).
- [36] V. N. Sathi, M. Srinivasan, P. K. Thiruvassagam and C. S. R. Murthy, "Novel Protocols to Mitigate Network Slice Topology Learning Attacks and Protect Privacy of Users' Service Access Behavior in Softwarized 5G Networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, pp. 2888–2906, November 2021. doi: [10.1109/tdsc.2020.2968885](https://doi.org/10.1109/tdsc.2020.2968885).
- [37] Y. Siriwardhana, P. Porambage, M. Liyanage, J. S. Walia, M. Matinmikko-Blue and M. Ylianttila, "Micro-Operator driven Local 5G Network Architecture for Industrial Internet," in *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, Marrakesh, 2019. doi: [10.1109/WCNC.2019.8885900](https://doi.org/10.1109/WCNC.2019.8885900).
- [38] E. J. D. Santos, R. D. Souza and J. L. Rebelatto, "Rate-Splitting Multiple Access for URLLC Uplink in Physical Layer Network Slicing With eMBB," *IEEE Access*, vol. 9, pp. 163178–163187, 2021. doi: [10.1109/ACCESS.2021.3134207](https://doi.org/10.1109/ACCESS.2021.3134207).
- [39] S. D'Oro, F. Restuccia, T. Melodia and S. Palazzo, "Low-Complexity Distributed Radio Access Network Slicing: Algorithms and Experimental Results," *IEEE/ACM Trans. Netw.*, vol. 26, pp. 2815–2828, December 2018. doi: [10.1109/TNET.2018.287896](https://doi.org/10.1109/TNET.2018.287896).

AI in disinformation detection

Julia Puczyńska | IDEAS NCBR Sp. z o.o., 69 Chmielna Street, 00-801 Warsaw, Poland, and IPPT PAN, Pawińskiego St. 5B; 02-106 Warsaw, Poland | ORCID: 0009-0009-5304-7092

Youcef Djenouri | IDEAS NCBR Sp. z o.o., 69 Chmielna Street, 00-801 Warsaw, Poland, and University of South-Eastern Norway (USN), Post Office Box 4, 3199 Borre, Norway | ORCID: 0000-0003-0135-7450

Abstract

The Russian Doppelganger campaign was a flop. It tried targeting European governments and institutions with fake news and cloned websites, but their measurable impact on real users — views, likes, or shares — was near zero [1]. However, as a part of continuous efforts to influence Western media, this campaign does contribute to changing the online discourse and normalising hate speech. The potential for harm from such attacks has proven to be even more extreme. Such threats require international efforts to identify and counter such campaigns effectively.

In this article, we consider the use of artificial intelligence (AI) in disinformation detection. The recent explosion of AI performance and popularity is a double-edged sword. On the one hand, AI makes generating fake news faster. On the other hand, it helps fight back; in fact, nowadays leveraging AI-driven techniques — such as Natural Language Processing (NLP), multimedia analysis, and network analysis — is crucial in the fight against fake news.

Our discussion is based on the DISARM Framework, a disinformation-focused counterpart to the MITRE ATT&CK® framework, designed to standardise disinformation-related terminology and analytical methods [2]. We focus particularly on a key tactic of disinformation that relies on overwhelming the target, apparent in many social engineering plots. Be it news or messages, the 21st century is overfilled

Received: 17.11.2024


Accepted: 20.12.2024

Published: 30.12.2024

Cite this article as:

Julia Puczyńska,
Youcef Djenouri, "AI in
disinformation detection,"
ACIG, vol. 3, no. 2,
2024, pp. 211–232. DOI:
10.60097/ACIG/200200

Corresponding author:

Julia Puczyńska, 69
Chmielna Street, 00-801
Warsaw, Poland, julia.
puczynska@ideas-ncbr.pl
 0009-0009-5304-7092

Copyright:

**Some rights reserved
(CC-BY):**

Julia Puczyńska,
Youcef Djenouri
Publisher NASK



with content, forcing people into constant stress, weakening their decision-making, and increasing their susceptibility to manipulation. We discuss the practical overview of disinformation detection. In this discussion, we include uncertainty quantification (UQ) as a groundbreaking tool to counteract this challenge (a solution introduced by Julia Puczyńska, Youcef Djenouri, Tomasz Pawel Michalak and Piotr Sankowski in 'Knowledge Base Monte Carlo for Uncertainty Quantification in Fake News Detection', mimeo, IDEAS NCBR, 2024). UQ enhances reliability, explainability, and adaptability in disinformation detection systems, as it enables estimation of model confidence.

Our framework demonstrates the potential of AI-driven systems to counteract disinformation through multimodal analysis and cross-platform collaboration while maintaining transparency and ethical integrity. We underscore the urgency of integrating UQ into fake news detection methodologies to address the rapid evolution of disinformation campaigns. The paper concludes by outlining future directions for developing scalable, transparent, and resilient systems to safeguard information integrity and societal trust in an increasingly digital age.

Keywords

disinformation, fake news, artificial intelligence, uncertainty quantification, social media

1. Introduction

Disinformation became a very popular topic after the 2016 US presidential election and again after Russia's 2022 invasion of Ukraine, and now artificial intelligence (AI)-powered technologies are raising the stakes even further. They're powering sophisticated disinformation campaigns, through, for example, Natural Language Processing (NLP) and generative AI models that help spread falsehoods at lightning speed [3].

Ironically, these same technologies provide innovative solutions to identify, analyse, and counteract disinformation. However, there's a gap between the tools researchers write about in their papers and the ones that actually get used. People on the frontlines of combating disinformation often do not know whether these solutions exist, cannot apply them, or cannot afford to integrate them into their work. This is why we believe it is time to bridge this gap. In this paper, we dig into how AI both enables and combats disinformation. We are using the 'Doppelganger' campaign as a base for a

case study, an example for the challenge, and a reminder of what's at stake.

We emphasise the need for a comprehensive approach to disinformation detection that combines technological innovation and ethical responsibility. This should include accessible and explainable AI-powered, uncertainty quantification (UQ)-based, robust fact-checking systems. We argue that the same technological advances fueling disinformation can and must be harnessed to safeguard the truth and rebuild societal trust.

1.1. Doppelganger Campaign

'Olaf Scholz has betrayed the German economy' says a bold headline, 'European Union will manage without Poland' – says another headline on the Polish Radio's site. Or do they? The Doppelganger campaign got its name for impersonating trusted media sources and spreading such disinformation. It is attributed to Russian influence operations and has been actively spreading propaganda in the United States, Germany, and Ukraine [1]. As of today (December 2024), the researchers from Recorded Future's Insikt Group are tracking over 2000 fake social media (SM) accounts associated with this campaign, which relies on impersonating news outlets and creating fake websites to disseminate false narratives. Key tactics include undermining Ukraine's political stability, military strength, and international alliances; promoting narratives of Germany's domestic decline; and exploiting the US political and social divisions ahead of the 2024 election. Notably, some content is likely generated using AI, reflecting an evolving approach to bypass detection and establish long-term influence networks.

The campaign has been linked to Russian companies Structura National Technologies and Social Design Agency, both sanctioned by the European Union (EU) and the United States for their involvement. These operations highlight the Kremlin's strategic use of disinformation in its broader information warfare, leveraging AI tools to scale propaganda efforts.

The campaign's attack flow, as illustrated in Figure 2, is focused on a singular goal, which is spreading content. These undertaken steps made it very persistent, despite the continuous efforts to mitigate its spread. However, as mentioned above, the campaign's reach is negligible compared to the resources it requires. It would seem that the sole purpose of these actions is the content's generation and not its appeal or its reach.

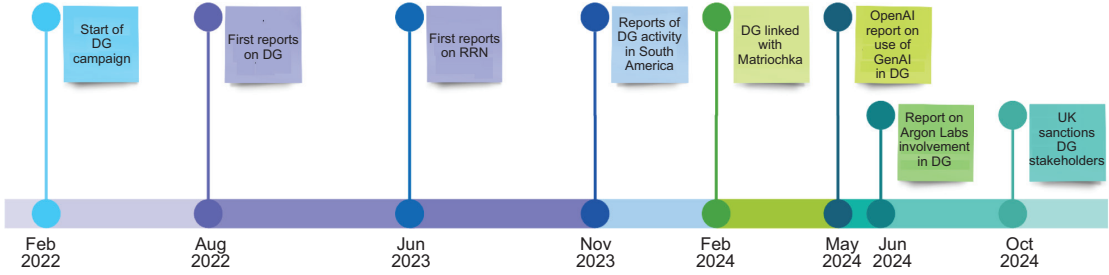


Figure 1. Timeline of reports regarding the Doppelganger (DG) campaign and the linked sub-campaigns.

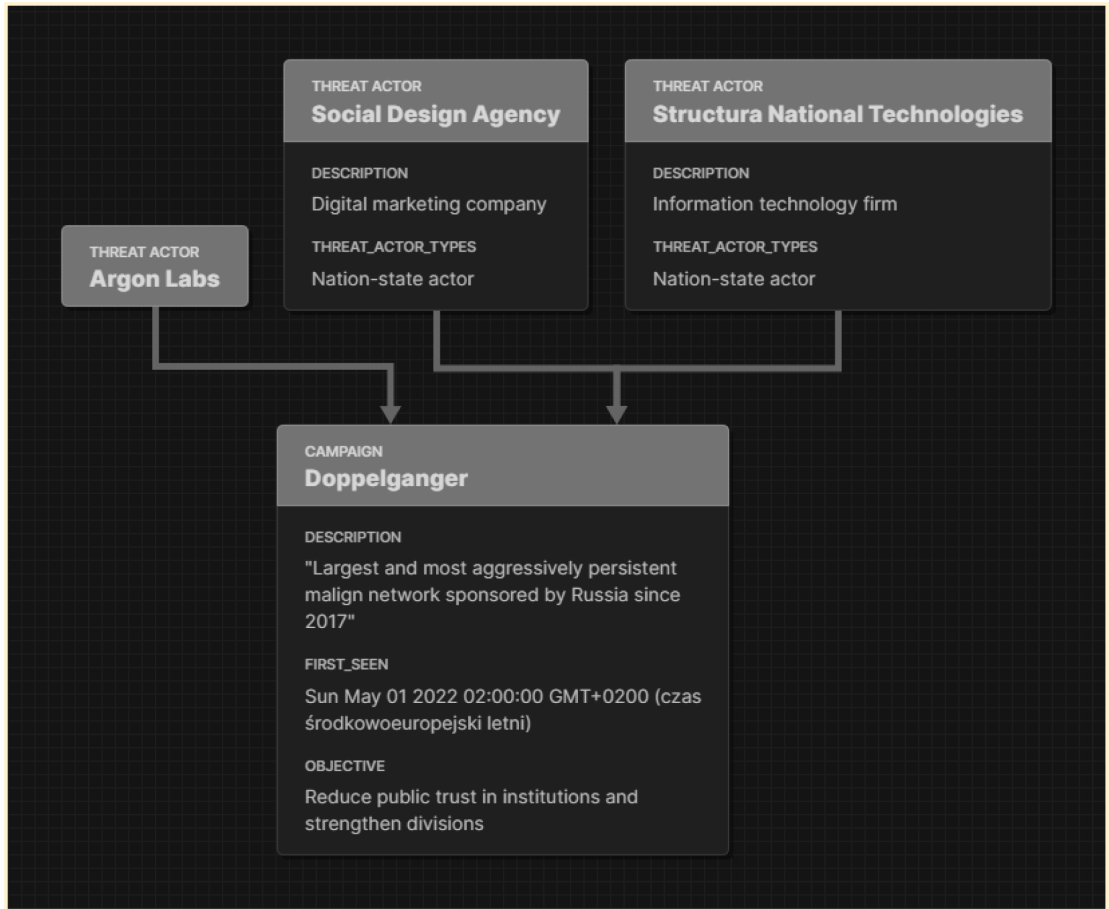


Figure 2. Doppelganger campaign-related threat actors.

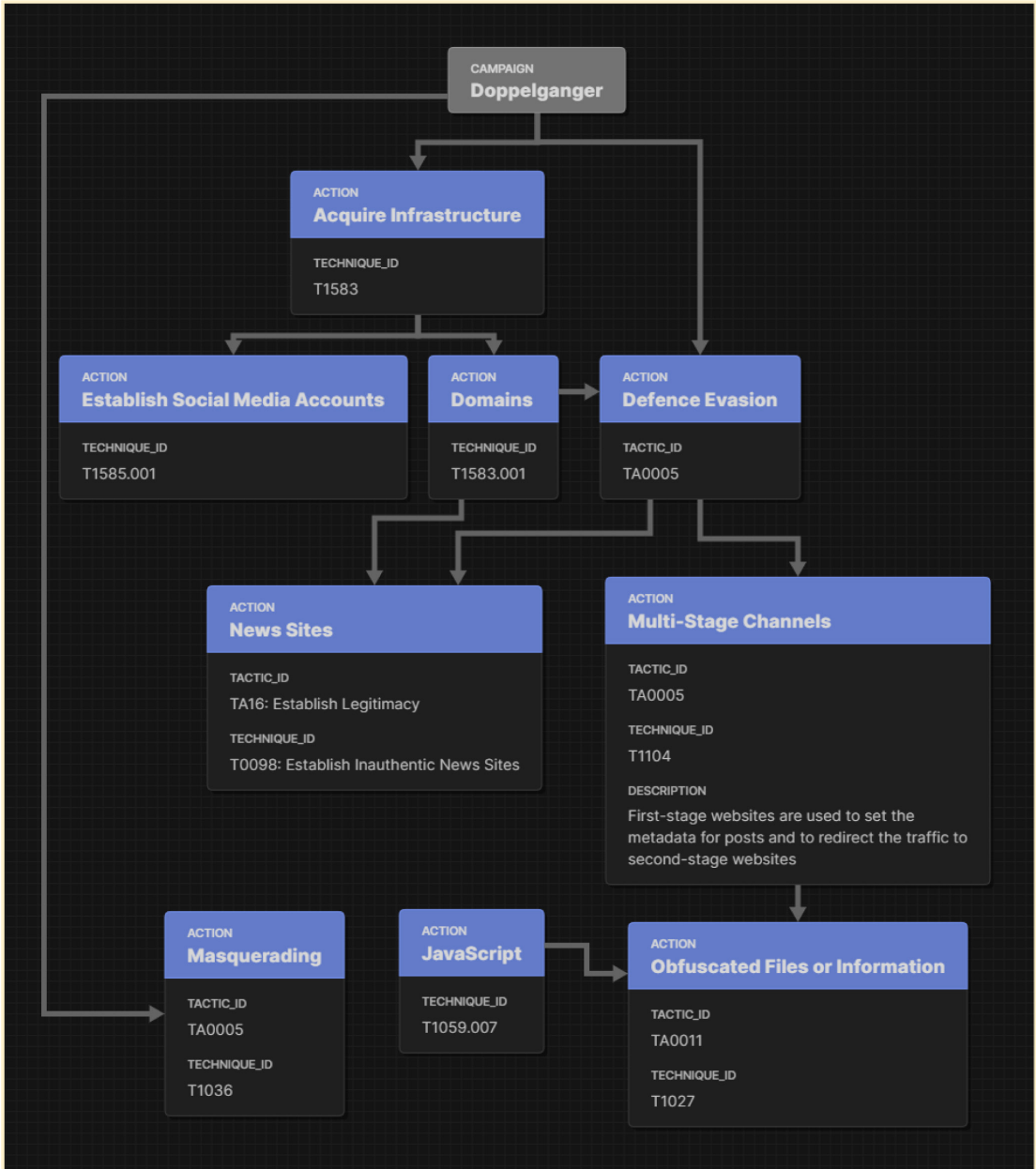


Figure 3. The campaign's attack flow – technical aspects.

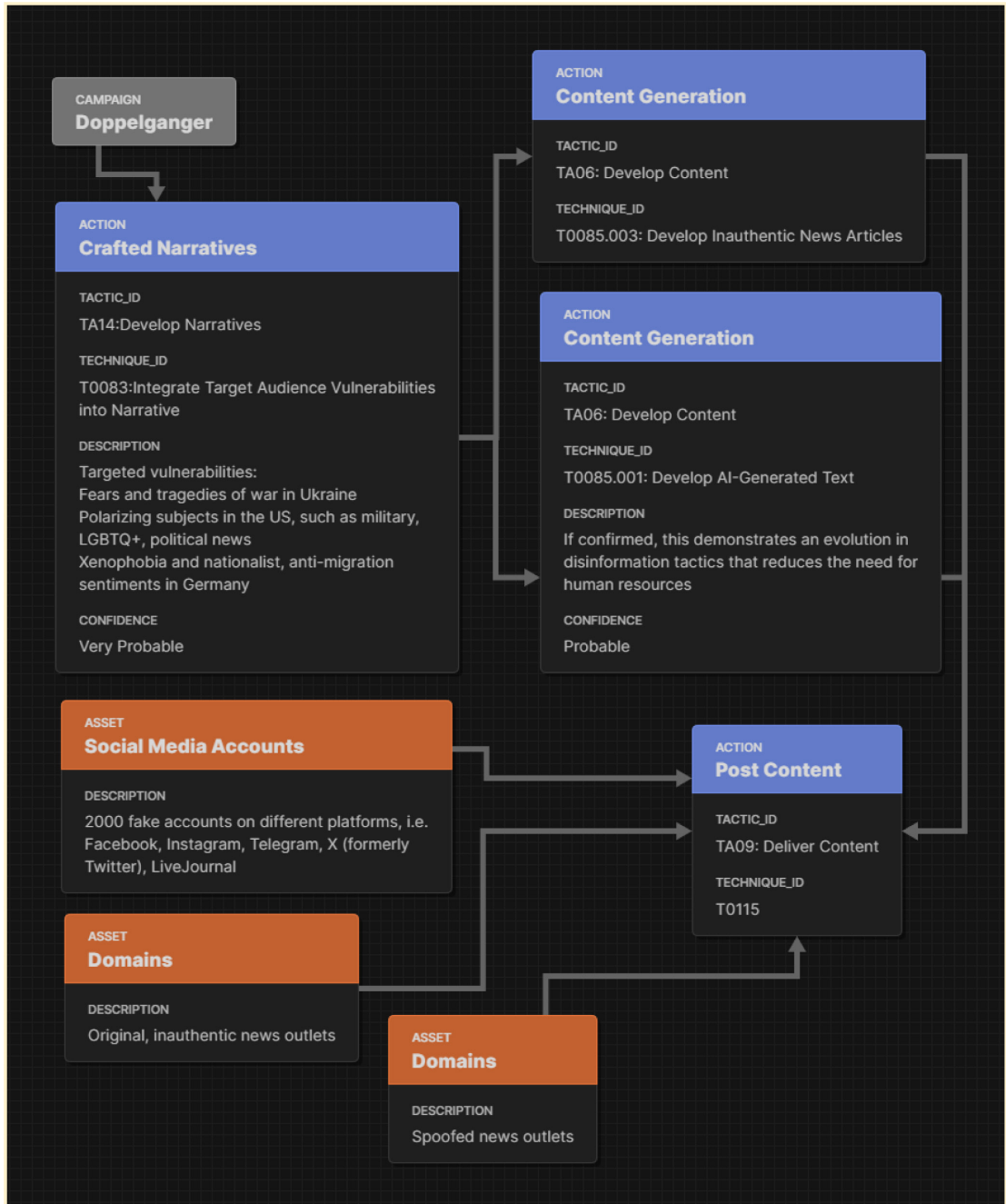


Figure 4. The campaign's attack flow – narrative aspects. The campaign lacked measures to adjust and read just content to the target audiences. While it is difficult to assess the true extent of threat actors' efforts to analyse and construct their messages, the reality still is that they were not well fitted to their audience.

1.2. Contributions

The objective of this paper is to widen the perspective on disinformation. Our contributions are as follows:

- We apply the DISARM Framework to the Doppelganger campaign analysis. This helps with cementing the framework's role in disinformation-related research.
- We discuss a practical, AI-based approach to current problems in disinformation detection.
- We highlight the role of UQ in disinformation detection as a solution to the problem of strained content moderation and fact-checking apparatus.

2. Background

Disinformation refers to deliberately false or misleading information created and shared with the intent to deceive or manipulate public opinion [4]. Unlike misinformation, which involves spreading incorrect or misleading information without malicious intent, disinformation is intentionally crafted to cause harm, confusion, or disruption. Fake news, a term frequently used in the context of SM, is a specific type of disinformation. It involves fabricated stories or media designed to resemble legitimate news, intending to deceive readers [5]. While fake news is always based on a lie, it often serves as a vehicle for spreading either disinformation or misinformation. The primary distinction between these terms lies in the intent and factuality, with disinformation being intentional and fake news always rooted in fabrication.

Recognition of these differences is crucial for developing effective strategies to combat harmful content and introduce appropriate consequences for its spread. Specifically, in our understanding unaware users share disinformation without intent to deceive, that is not misinformation, because the content itself is being crafted and originally shared in order to manipulate the recipient. Therefore, while simply detecting harmful and misleading content usually does not include detection of intention (which is difficult to establish), we choose to keep this definition in order to retain the induced accountability for both its creation and spread.

2.1. Artificial Intelligence

Artificial Intelligence is a branch of computer science that aims to develop systems capable of performing tasks that typically require human intelligence, such as learning, reasoning,

problem-solving, and perception [6]. In the realm of disinformation, AI plays a dual role. On the one hand, it is used to create misleading content, such as deepfakes [7] and, on the other, it is utilised to combat disinformation through various detection and verification systems [8]. AI technologies, particularly NLP [9] and Large Language Models (LLMs) [10], are instrumental in both creation and identification of false narratives. NLP, a subfield of AI, focuses on enabling machines to understand, interpret, and generate human language. It plays a critical role in disinformation detection, as it can analyse patterns in online conversations, identify manipulated text, and track emerging trends. For instance, sentiment analysis techniques in NLP can identify manipulative language, often present in disinformation, by classifying text as having a positive, negative, or neutral sentiment [11].

The sentiment score can be as simple as a mean average of sentiment value associated with each word in a piece of text; that is, it can be calculated using a simple formula:

$$S = \frac{\sum_{i=1}^N \text{score}(w_i)}{N}, \quad (1)$$

where w_i represents individual words in the text, and $\text{score}(w_i)$ is the sentiment score for each word, which is typically drawn from a pre-defined lexicon. The value of N is the number of words in the document.

2.2. Large Language Models

Large Language Models, such as Open AI's GPT models and Google's Bard, are trained on vast datasets to generate and understand text. These models contribute to the creation of sophisticated fake narratives by bots and are also used to counter disinformation by performing advanced text analysis, summarisation, and verification tasks. LLMs rely on neural networks that process vast amounts of textual data and learn the underlying patterns of language. For example, GPT models [12, 13] use transformer architecture, and the model's responses are based on both input and a set of parameters, which are defined during training. The transformer model uses self-attention mechanisms to weigh the importance of each word relative to others, enabling it to capture syntactic and semantic relationships in language and to generate a coherent text.

This process can be represented using the following transformation:

$$y = f(x, \theta), \quad (2)$$

where X is the input sequence (text data), θ represents the model's parameters, and y is the predicted output (e.g., the next word in the sequence).

2.3. Uncertainty Quantification

It is a mathematical framework designed to assess the uncertainty inherent in model predictions [14]. By identifying areas where predictions are uncertain, UQ provides confidence levels for specific outcomes, allowing for more informed decision-making. In the context of AI systems used for disinformation detection, UQ can help improve the robustness of models by quantifying their reliability. Simply put, where a model can classify content as disinformation or not, UQ returns the certainty of such classification, so how sure we are that this response is accurate.

Statistical inference based on a single data point, for example, an article, requires artificial multiplication of data. The article can then be assessed as false with a 70% confidence – because in 70% of these multiplied cases the article has been deemed false. One common approach in UQ is the use of Bayesian methods [15], which infer distributions over model parameters. This allows for a more probabilistic interpretation of model predictions, rather than providing deterministic outputs. For instance, if we have a model that predicts the likelihood \hat{y} of a claim being false, the Bayesian approach provides a distribution over the prediction:

$$p(\hat{y} | x) = \int p(\hat{y} | \theta, X) p(\theta | X) d\theta, \quad (3)$$

where θ represents the model parameters, and $p(\theta | X)$ represents the updated probability distribution of the model's parameters after

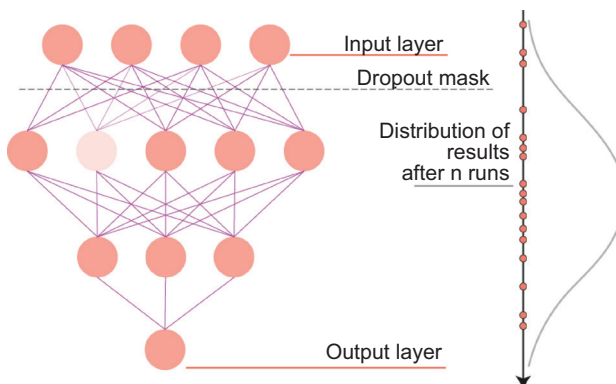


Figure 5. Monte Carlo Dropout.

incorporating the observed data X . This distribution represents the uncertainty in the model's predictions, allowing us to quantify how confident the model is about its conclusions.

A practical example of UQ in disinformation detection can be found in Monte Carlo Dropout [16], a method that estimates the uncertainty by applying dropout during inference. Dropout is a technique typically used during training to prevent over fitting, where certain neurons in the neural network are randomly 'dropped' or ignored during each forward pass. To quantify uncertainty, Monte Carlo Dropout keeps the dropout layers active during inference. The final prediction \hat{y} is made by averaging multiple forward passes, each with different random neurons omitted, producing a distribution of predictions:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N f(X, \theta_i), \quad (4)$$

where θ_i are the parameters of the model after each forward pass with different dropout configurations, and N is the number of forward passes. The variance across these predictions provides an estimate of the model's uncertainty.

Another method used in UQ is Deep Ensembles [17], where multiple models are trained on the same dataset and their predictions are aggregated to estimate uncertainty. This approach captures the range of possible outcomes by training different models, each with slightly varied parameters, and combining their predictions. The uncertainty can then be calculated as the variance between the predictions of ensemble models:

$$U(\hat{y}) = \frac{1}{M} \sum_{j=1}^M \hat{y}_j - \hat{y}_{avg}, \quad (5)$$

where \hat{y}_j is the prediction from the j -th model in the ensemble, \hat{y}_{avg} is the average prediction across all models, and M is the number of models in the ensemble.

3. Practice

As mentioned, there is a disparity between the current state-of-the-art solutions in theoretical works and the solutions actually employed by some of those that take on the responsibility to fact check viral news. Therefore, we decided to describe in detail the reality of fact-checking.

3.1. Sources of Disinformation

Disinformation is similar to scams: both are based on influencing people and exploiting their vulnerability at the moment of contact with manipulative content. The endgoals may differ, but plenty of methods stay the same: overwhelming, inspiring fear, and impersonating trusted figures and sources. The spread of disinformation is incredibly complicated because it includes different social media platforms, TV, newspapers both online and offline, advertisements, and simply word-of-mouth [2]. Depending on the demographics, different sources matter more and less, but social media is increasingly significant.

One of the reasons for social media's popularity amongst security researchers is that regulations cannot keep up with the underlying technology. While the literature might already exist for plenty of possible threats related to social media, the public is often not informed and equipped well enough to recognise and appropriately react to them. It is worth noting that plenty of people base their knowledge about current events on social media. Polish IBIMS (Instytut Badań Internetu i Mediów Społecznościowych) and IBRIS report investigated the percentage of users who draw information from the Internet but differentiated between online news outlets (60% of respondents) and social media (38.8%) [18]. However, it is worth noting that users often access articles through links to news outlets on social media. Therefore, they are still subjected to, for example, biased selection of content by the social media algorithms. Interestingly, the Doppelganger campaign's fake news websites, which posed as trusted news outlets, could only be accessed through links in sponsored content posted on Facebook and X (formerly Twitter).

3.2. Fact-Checking

Efforts to combat disinformation involve a combination of strategies from governments, non-governmental organisations (NGOs), and social media companies. Each entity approaches the problem from different angles, leveraging its unique capabilities and areas of influence.

3.3. Government

Governments often focus on regulatory measures, public awareness campaigns, and collaboration with organisations to mitigate the spread of disinformation. For example, the EU's Digital Services Act (DSA) holds platforms accountable for harmful

content, including disinformation. Dedicated bodies to monitor and counteract disinformation include the US Cybersecurity and Infrastructure Security Agency (CISA, at <https://www.cisa.gov/>) or the European External Action Services East StratCom Task Force, which runs the EUvsDisinfo project – a database of articles and media considered to be disinformative (at <https://euvsdisinfo.eu/>). Such organisations create and implement their solutions, which have proven useful in France against the Doppelganger campaign. Their Service for Vigilance and Protection against Foreign Digital Interference (VIGINUM) agency, subject to Secretariat-General for National Defence and Security (fr. *Secrétariat général de la défense et de la sécurité nationale*, SGDSN), detected imitations of four French media outlets [19]. The organisation reported its findings on the RRN (*rrussianews*), an anonymous newsmedia organisation behind these fake websites. The RRN serves as a content repository for Doppelganger [1].

3.4. Non-Governmental Organisations

Non-governmental organisations focus on research, education, and advocacy to combat disinformation while supporting free speech and human rights. NGOs, like FactCheck.org, PolitiFact, and the International Fact-Checking Network (IFCN), identify and debunk disinformation and often partner with social media companies to label or flag false content (<https://www.poynter.org/ifcn/>) (Accessed: Nov. 18, 2024). This helps to create data sets for training disinformation detection systems [20]. Such organisations also develop programmes to improve critical thinking skills and media literacy among the public (<https://geremek.pl/program/cyfrowa-akademia-walki-z-dezinformacja/>) (Accessed: Nov. 18, 2024).

Input through NGOs is invaluable. Their impartial nature makes them the much-needed judges of the system's efficacy and equity. However, that also means they are highly dependent on donations, which often lead to underfunding and understaffing. Government and corporate funding helps solve this problem. In turn, it largely affects the impartiality of these organisations.

3.5. Social Media Companies

Social media platforms, as primary vectors for disinformation, focus on improving content moderation, transparency, and user awareness. Strategies include content moderation using AI and human moderators to detect and label or remove disinformation.

Platforms actively suspend fake or bot accounts spreading disinformation. In some cases, they remove coordinated inauthentic behaviour, as seen in campaigns linked to state-sponsored actors. Meta's Ad Transparency Tool is an example of providing access to information about political ads and their funding. It is worth noting that plenty of their efforts are legally required, for example, the Ad Transparency Tool or counteracting the spread of disinformation and reporting the results of their efforts. Without the regulation, these platforms wouldn't have the incentive to invest in countering the spread of disinformation; which became especially clear when Meta resigned from fact-checking programs in favor of an X-style 'community notes'.

4. Challenges for Disinformation Detection Framework

Fact-checking faces numerous challenges in the digital age. These obstacles can be broadly categorised into technical, operational, and societal domains, each presenting unique complexities that must be addressed for effective disinformation detection and mitigation.

4.1. Technical challenges

Volume and velocity: The digital ecosystem generates daily an overwhelming volume of content, ranging from social media posts and news articles to multimedia content. The rapid pace at which disinformation spreads often outpaces fact-checking efforts. Viral misinformation can reach millions within hours, while corrections, even when issued, struggle to achieve similar penetration. For example, during crises or high-profile events, false narratives dominate public discourse long before accurate information is disseminated. This imbalance underscores the need for scalable, automated tools capable of processing and verifying large quantities of data in real time.

Lack of datasets: Available datasets for disinformation detection include: FakeNewsNet, LIAR, ISOT FakeNews Dataset and WEIBO. However, more datasets are needed: firstly, there is a need for diverse datasets, including platform-specific and language-specific data. Nuances and contexts present in different cultures, platforms, and modalities are underrepresented. Existing datasets focus predominantly on text, leaving a gap in multimodal detection capabilities, and limiting the applicability and usefulness of fake news detection systems. Secondly, new topics and forms

of disinformation arise during global events (e.g., pandemics, elections, wars). Dynamic and up-to-date datasets are crucial to address evolving challenges. Since platforms like X, Instagram, and Facebook are not supported by any fact-checking programs, the access to data is limited even more.

Generative AI: Recent advancements in AI, particularly in generative technologies, have exacerbated the challenge. Tools like large language models and generative adversarial networks (GANs) are now capable of creating highly convincing fake content [21, 22, 23]. Deepfake videos can depict public figures engaging in fabricated acts, while AI-generated articles mimic credible news sources with alarming accuracy. The sophistication of such content makes it difficult for both humans and existing automated tools to discern authenticity, requiring the development of advanced detection algorithms tailored to generative outputs.

Multimodal disinformation: Disinformation campaigns increasingly utilise multimodal formats, blending text, images, and videos to enhance believability and engagement [24]. For instance, a false claim might be accompanied by a doctored image or a video with altered context, creating a layered narrative that appears credible. Detecting and analysing such multimodal disinformation demands cross-modal AI systems capable of correlating information across different formats – a complex and resource-intensive task.

4.2. Operational Challenges

Cross-platform propagation: Disinformation effortlessly migrates across platforms, exploiting the lack of coordinated detection mechanisms between social media, messaging apps, and traditional news outlets. A false narrative might originate on one platform, such as a tweet, and subsequently be amplified on others, including Facebook, Instagram, or WhatsApp. This fragmented ecosystem complicates detection efforts, as each platform employs varying policies, tools, and capabilities to address disinformation. Building interoperable solutions and fostering collaboration among platforms is critical but remains an unresolved challenge.

Language and cultural nuances: Disinformation often leverages specific linguistic and cultural contexts to increase its impact. A narrative tailored for one region may exploit local events, historical tensions, or societal biases, making it challenging to detect using generalised tools. Furthermore, many fact-checking systems and datasets are optimised for dominant languages like English, leaving significant

gaps in coverage for regional languages and dialects. Effective detection requires a nuanced understanding of cultural context and linguistic subtleties, necessitating localised datasets and AI models.

4.3. Societal Challenges

Polarisation and bias: In politically polarised environments, fact-checking is often perceived as an extension of one ideological viewpoint, undermining its credibility. Disinformation campaigns exploit these divisions, framing corrections as biased attempts to suppress dissenting opinions. This skepticism is further fueled by bad actors who discredit fact-checkers and promote narratives of censorship.

Overcoming this challenge requires transparent methodologies, diverse fact-checking teams, and the inclusion of multiple perspectives in verification processes to build public trust.

Trust deficits: A growing distrust in institutions, including media organisations and fact-checking bodies, significantly hampers efforts to combat disinformation. In many cases, people are more likely to trust information shared within their social or ideological circles than corrections issued by external entities. Addressing this trust deficit involves not only improving the accuracy and transparency of fact-checking efforts but also engaging communities directly to foster grassroots awareness and resilience against disinformation.

To overcome these challenges, a multi-pronged approach is required. Technical advancements must prioritise scalability and multimodal capabilities. Operational strategies should emphasise cross-platform collaboration and localised solutions. On the societal front, rebuilding trust through transparency, community engagement, and education is imperative. These measures, when integrated into a cohesive framework, can enhance the effectiveness of fact-checking efforts in the digital age.

5. Methodology and Tool Set

The dynamic and multifaceted nature of disinformation necessitates a diverse arsenal of tools that automate and enhance the detection process. These tools leverage cutting-edge AI, statistical techniques, and domain-specific expertise to identify, verify, and counter disinformation [25]. They can be broadly categorised into text analysis tools, multimedia analysis tools, and network analysis tools, each addressing specific challenges in the fact-checking landscape.

Text tools analysis: NLP techniques are pivotal in identifying and countering textual disinformation by analysing the tone, intent, and content of the text. Sentiment analysis helps flag emotionally charged or manipulative language often used in disinformation, such as fear-mongering or sensationalism designed to provoke rapid sharing without scrutiny. Entity recognition, another critical NLP capability, extracts and categorises names, organisations, and locations within a text, enabling cross-referencing with trusted databases to spot inconsistencies or fabrications. Claim matching, meanwhile, identifies recurring patterns or exact matches of previously debunked statements, aiding in the rapid recognition of recycled disinformation narratives. Beyond these, advanced language models like GPT, BERT, and T5 enhance the process by retrieving and cross-referencing documents from credible sources to verify claims [26]. Where simple sentiment analysis may fail, these models excel in understanding complex linguistic nuances, such as sarcasm or context-dependent meanings, which are often employed in sophisticated disinformation. Furthermore, integrating such models into automated fact-checking pipelines, supported by structured datasets, accelerates the generation of fact-checking reports for emerging claims, providing a scalable and efficient approach to combating textual disinformation [27].

Multimedia tools analysis: The rise of multimedia disinformation has necessitated the development of specialised tools for analysing and detecting manipulated visual content, from altered images to synthetic videos [28]. Image forensics plays a crucial role by examining metadata – such as timestamps, geolocation, and camera settings – to uncover inconsistencies indicative of tampering [29]. Algorithms also detect visual artifacts like irregular pixel patterns, lighting mismatches, or compression anomalies, which often result from editing processes. Additionally, reverse image search techniques allow cross-referencing of suspect visuals with existing databases to identify duplicates or modifications. Similarly, video analysis tools tackle the challenges posed by deepfakes and spliced footage. Biometric inconsistencies, such as unnatural blinking or misaligned facial expressions, are flagged using deepfake detection algorithms, including those that analyse generative model fingerprints imperceptible to humans. Temporal analysis further aids detection by identifying irregularities in motion, lighting, or audio synchronisation, which often signal manipulation. Advanced scene reconstruction techniques complement these efforts by contextualising video content, enabling evaluators to determine whether depicted events genuinely align with the associated narrative. Together, these tools

form a comprehensive framework for combating multimedia-based disinformation.

Network tools analysis: Disinformation campaigns often leverage complex dissemination networks to amplify their reach and legitimacy, necessitating robust network analysis tools for effective detection and mitigation. Propagation mapping is a critical technique in this context, allowing researchers to track the evolution and spread of disinformation narratives across social platforms [30]. By identifying key actors, influential hashtags, and clusters responsible for amplifying false information, these tools enable targeted interventions. Algorithms that detect influential nodes within the network – individuals or groups with a disproportionate impact on disinformation dissemination – are particularly valuable in disrupting these campaigns. Temporal dynamic analysis further strengthens this approach by examining the timing and frequency of posts to identify patterns indicative of coordinated campaigns, such as those orchestrated by bot networks or state-sponsored entities.

Bot detection forms another essential component of network analysis, addressing the role of automated accounts in disseminating disinformation. The behavioural analysis identifies suspicious patterns, such as excessive posting frequency, identical content shared across multiple accounts, or activity during improbable hours, all of which suggest automation. Network-specific features, including low engagement rates, clustering within particular communities, or repeated interactions with known disinformation agents, further assist in distinguishing bots from human users. Machine learning models trained on diverse datasets enhance this process, classifying accounts based on multidimensional behavioural characteristics. Together, these tools provide a comprehensive approach to mapping, analysing, and ultimately disrupting the networks that propagate disinformation.

Other tools: We believe that there is a need for reliable and accessible fact-checking tools that can be used by both specialists and general public. These include web plug-ins, news apps, and dedicated SM profiles; all of these should focus on increasing the users' ability to determine what is trustworthy. The 'Ground News' app, which aims to provide informative news headlines and insight into the bias of reported news, serves as a great example of what is wildly needed. Today's users are overwhelmed with content. Anything that helps with limiting the quantity of content they receive, without

jeopardising its quantity and the users' choice over what they can get access to, is of value.

6. Vision for the Doppelganger

The Doppelganger campaign succeeds mainly in the sheer amount of content, created and/or generated by AI – what it lacks in likes, views, and shares, it makes up for in scale and persistence. While it may seem like a waste of resources, we think that the actual lesson that needs to be learned from it is that it would not take much for this campaign to be significantly more successful. Had these articles and their content caught on and spread among actual users, the mitigation couldn't have been limited to the continuous blocking of websites and fake accounts. Once real users would have been involved, their accounts would often not be blocked just because they shared disinformative content and the content itself might not be blocked nor marked as untrue. Such infrastructure as the one created for the sake of the Doppelganger campaign would keep providing new articles and links for these users, overwhelming even more our already strained system.

6.1. Uncertainty Quantification

Uncertainty quantification presents a transformative opportunity to enhance the robustness and reliability of fake news detection systems, particularly in the context of complex disinformation campaigns like Doppelganger. This campaign, which relied on cloned websites and targeted social media manipulation, demonstrates the challenges of distinguishing fabricated narratives from legitimate content. Traditional detection models often provide binary classifications, lacking the nuanced confidence metrics needed to guide critical decisions. Integrating UQ into these systems can address this limitation by estimating the reliability of predictions. For example, when analysing cloned content, UQ can pinpoint regions of high uncertainty, prompting additional human verification. Similarly, in cross-platform disinformation campaigns, where the context and format of narratives can vary, UQ can identify instances of low-confidence classifications. This capability ensures that questionable results are flagged for further scrutiny, reducing the risk of false positives or missed threats.

In addition to bolstering detection accuracy, UQ enhances the adaptability and transparency of fake news detection frameworks. Disinformation campaigns like Doppelganger evolve rapidly, with adversaries employing novel tactics to evade detection. UQ enables

systems to dynamically recalibrate their predictions by identifying areas where the model lacks sufficient training data or encounters new patterns. This adaptive capability ensures resilience against the evolving tactics of disinformation actors. Furthermore, the integration of UQ fosters greater transparency, particularly in politically sensitive contexts. By providing explanations alongside confidence metrics, UQ empowers stakeholders – such as fact-checkers, policymakers, and the public – to better understand and trust the decisions made by AI-driven detection systems. This transparency is critical in countering skepticism and ensuring that automated systems are perceived as reliable partners in combating disinformation.

Looking forward, UQ can play a pivotal role in improving the efficiency of resource allocation and the overall scalability of fake news detection efforts. Disinformation campaigns operate on a massive scale, often overwhelming human fact-checkers and investigative teams. UQ facilitates the prioritisation of high-risk cases by flagging predictions with elevated uncertainty for manual review. This targeted approach allows human resources to focus on the most critical and ambiguous cases, improving the efficiency of detection efforts. Furthermore, UQ strengthens defences against adversarial tactics, such as subtle content modifications that seek to exploit detection system vulnerabilities. By identifying instances of high uncertainty – often indicative of adversarial interference – UQ provides an early warning system for emerging threats. Finally, as cross-platform disinformation becomes more prevalent, the standardisation of UQ protocols enables seamless collaboration between platforms, fostering trust towards automated fact-checking and enabling coordinated responses to campaigns like Doppelganger. Together, these advancements position UQ as a cornerstone of future efforts to safeguard information integrity and societal trust.

7. Conclusions

In conclusion, the fight against disinformation, exemplified by campaigns, like Doppelganger, presents a growing challenge in the digital age. The dual role of AI in enabling and mitigating disinformation underscores the complexity of addressing this issue effectively. This paper has outlined a comprehensive framework for disinformation detection, emphasising the importance of integrating advanced AI techniques, such as NLP, multimedia analysis, and network analysis, into the detection process. Moreover, it has discussed UQ as a critical innovation, offering enhanced reliability

and interpretability for AI-driven detection systems. UQ not only improves confidence in predictions but also provides valuable insights that can guide human intervention, prioritise resources, and ensure the system remains adaptable to emerging disinformation tactics. As disinformation campaigns continue evolving in sophistication and scale, the need for adaptive, transparent, and collaborative detection mechanisms becomes increasingly urgent. This framework offers a promising direction for developing systems that can not only identify false narratives across multiple platforms but also respond to them in a way that is both efficient and ethically responsible. Moving forward, future research should focus on refining UQ techniques, improving cross-platform collaboration, and developing scalable solutions that can handle the ever-increasing volume and velocity of disinformation. By harnessing the full potential of AI and UQ, we can build a more resilient and trustworthy information ecosystem, safeguarding truth and societal trust in an increasingly complex digital world.

8. Acknowledgements

The authors used LLMs for editing and polishing the author-written version of the text, replacing about 15% of the original text.

References

- [1] Insikt Group (2023). Obfuscation and AI content in the Russian influence network “doppelgänger” signals evolving tactics [Online]. Available: <https://www.recordedfuture.com/research/russian-influence-network-doppelgangers-ai-content-tactics> [Accessed: Nov. 18, 2024].
- [2] DISARM Foundation (2022). DISARM disinformation TTP (tactics, techniques and procedures) framework [Online]. Available: <https://github.com/DISARMFoundation/DISARMframeworks> [Accessed: Nov. 18, 2024].
- [3] J. Puczyńska, M. Podhajski, K. Wojtasik T. P. Michalak. *Large language models in jihadist terrorism and crimes*. Warsaw: Agencja Bezpieczeństwa Wewnętrznego (Internal Security Agency), 2024, 351 p.
- [4] S. Abdali, S. Shaham, B. Krishnamachari. *Multi-modal misinformation detection: Approaches, challenges and opportunities*. New York, NY: ACM Computing Surveys, 2022.
- [5] J. Alghamdi, S. Luo, Y. Lin, “A comprehensive survey on machine learning approaches for fake news detection,” *Multimedia Tools and Applications*, vol. 83, no. 17, pp. 51009–51067, 2024, doi: [10.1007/s11042-023-17470-8](https://doi.org/10.1007/s11042-023-17470-8).
- [6] Y. Xu, X. Liu, X. Cao, C. Huang, E. Liu, S. Qian, ... J. Zhang, “Artificial intelligence: A powerful paradigm for scientific research,” *The Innovation*, vol. 2, no. 4, p. 100179, 2021, doi: [10.1016/j.xinn.2021.100179](https://doi.org/10.1016/j.xinn.2021.100179).

- [7] Á. F. Gambín, A. Yazidi, A. Vasilakos, H. Haugerud, Y. Djenouri, "Deepfakes: Current and future trends," *Artificial Intelligence Review*, vol. 57, no. 3, p. 64, 2024, doi: [10.1007/s10462-023-10679-x](https://doi.org/10.1007/s10462-023-10679-x).
- [8] V. U. Gongane, M. V. Munot, A. D. Anuse, "A survey of explainable ai techniques for detection of fake news and hate speech on social media platforms," *Journal of Computational Social Science*, vol. 7, pp. 1–37, 2024, doi: [10.1007/s42001-024-00248-9](https://doi.org/10.1007/s42001-024-00248-9).
- [9] G. G. Devarajan, S. M. Nagarajan, S. I. Amanullah, S. S. A. Mary, A. K. Bashir, "AI-assisted deep NLP-based approach for prediction of fake news from social media users," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 4975–4985, Aug. 2024, doi: [10.1109/TCSS.2023.3259480](https://doi.org/10.1109/TCSS.2023.3259480).
- [10] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, P. Qi, "Bad actor, good advisor: Exploring the role of large language models in fake news detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 20, pp. 22105–22113, 2024, doi: [10.48550/arXiv.2309.12247](https://doi.org/10.48550/arXiv.2309.12247).
- [11] J. Li, L. Xiao, "Multi-emotion recognition using multi-EmoBERT and emotion analysis in fake news," in *Proceedings of the 15th ACM web science conference 2023 (WebSci '23)*. Association for Computing Machinery, New York, NY, USA, 128–135. doi: [10.1145/3578503.3583595](https://doi.org/10.1145/3578503.3583595).
- [12] V. Alto, *Modern generative AI with ChatGPT and OpenAI models: Leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4*. Birmingham: Packt Publishing, 2023.
- [13] K. S. Kalyan, "A survey of GPT-3 family large language models including ChatGPT and GPT-4," *Natural Language Processing Journal*, vol. 6, p. 100048, 2023. doi: [10.1016/j.nlp.2023.100048](https://doi.org/10.1016/j.nlp.2023.100048).
- [14] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, S. Nahavandi, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.
- [15] G. Franchi, A. Bursuc, E. Aldea, S. Dubuisson, I. Bloch, "Encoding the latent posterior of Bayesian neural networks for uncertainty quantification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2027–2040, April 2024, doi: [10.1109/TPAMI.2023.3328829](https://doi.org/10.1109/TPAMI.2023.3328829).
- [16] D. Bethell, S. Gerasimou, R. Calinescu, "Robust uncertainty quantification using conformalised Monte Carlo prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, pp. 20939–20948, 2024. doi: [10.1609/aaai.v38i19.30084](https://doi.org/10.1609/aaai.v38i19.30084).
- [17] R. Rahaman, "Uncertainty quantification and deep ensembles," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20063–20075, 2021.
- [18] IBIMS & IBRIS (2022). Skąd polacy czerpią informacje? [Online]. Available: <https://ibims.pl/wp-content/uploads/2021/01/Raport-IBIMS-IBRIS-Zrodla-informacji-Polakow.pdf> [Accessed: Nov. 18, 2024].
- [19] Service for Vigilance and Protection against Foreign Digital Interference (VIGINUM) (2023). RRN: A complex and persistent information manipulation campaign [Online]. Available: https://www.sgdsn.gouv.fr/files/files/Publications/20230719_NP_VIGINUM_RAPPORT-CAMPAGNE-RRN_EN.pdf [Accessed: Nov. 18, 2024].

- [20] K. Shu, G. Zheng, Y. Li, S. Mukherjee, A. H. Awadallah, S. Ruston, H. Liu, "Leveraging multi-source weak social supervision for early detection of fake news," arXiv preprint arXiv:2004.01732, 2020, doi: [10.48550/arXiv.2004.01732](https://doi.org/10.48550/arXiv.2004.01732).
- [21] R. Raman, V. K. Nair, P. Nedungadi, A. K. Sahu, R. Kowalski, S. Ramanathan, K. Achuthan, "Fake news research trends, linkages to generative artificial intelligence and sustainable development goals," *Heliyon*, vol. 10, no. 3, p. e24727, 2024, doi: [10.1016/j.heliyon.2024.e24727](https://doi.org/10.1016/j.heliyon.2024.e24727).
- [22] A. Bashardoust, S. Feuerriegel, Y. R. Shrestha, "Comparing the willingness to share for human-generated vs. AI-generated fake news," *Proceedings of the ACM on Human-Computer Interaction*, vol. 8, no. CSCW2, pp. 1–21, 2024, doi: [10.1145/3687028](https://doi.org/10.1145/3687028).
- [23] D. Xu, S. Fan, M. Kankanhalli, "Combating misinformation in the era of generative AI models," in *Proceedings of the 31st ACM international conference on multimedia*, Association for Computing Machinery, New York, NY, USA, 9291–9298, 2023, doi: [10.1145/3581783.3612704](https://doi.org/10.1145/3581783.3612704).
- [24] S. Tufchi, A. Yadav, T. Ahmed, "A comprehensive survey of multimodal fake news detection techniques: Advances, challenges, and opportunities," *International Journal of Multimedia Information Retrieval*, vol. 12, no. 2, p. 28, 2023, doi: [10.1007/s13735-023-00296-3](https://doi.org/10.1007/s13735-023-00296-3).
- [25] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, F. Menczer, "Detecting and tracking political abuse in social media," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, pp. 297–304, 2011, doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423).
- [26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the Association for computational linguistics: Human language technologies*, vol. 1, p. 2, pp. 4171–4186. Association for Computational Linguistics, Kerrville, TX 78028, USA, 2019, doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423).
- [27] S. Bansal, N. S. Singh, S. S. Dar, N. Kumar, "MMCFND: Multimodal multilingual caption-aware fake news detection for low-resource indic languages," arXiv preprint, arXiv:2410.10407, 2024, doi: [10.48550/arXiv.2410.10407](https://doi.org/10.48550/arXiv.2410.10407).
- [28] F. Marra, D. Gragnaniello, D. Cozzolino, L. Verdoliva, "Detection of generated fake images over social networks," in *Proceedings of the 2018 IEEE conference on multimedia information processing and retrieval (MIPR)* pp. 384–389, 2018, New York, NY: IEEE. doi: [10.1109/MIPR.2018.00084](https://doi.org/10.1109/MIPR.2018.00084).
- [29] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020, doi: [10.1109/JSTSP.2020.3002101](https://doi.org/10.1109/JSTSP.2020.3002101).
- [30] L. Vargas, P. Emami, P. Traynor, "On the detection of disinformation campaign activity with network analysis," in *Proceedings of the 2020 ACM SIGSAC conference on cloud computing security workshop*, pp. 133–146, ACM. New York, NY, 2020. doi: [10.1145/3411495.3421360](https://doi.org/10.1145/3411495.3421360).

Exploiting Human Trust in Cybersecurity: Which Trust Development Process is Predominant in Phishing Attacks?

Morice Daudi | Computing Science Studies, Mzumbe University, Tanzania |
ORCID: 0000-0001-7907-427X

Abstract

Humans live in an interconnected world that is increasingly featured with virtual interactions in cyberspace. That world has raised cybersecurity concerns, particularly on exploiting human trust through various means, such as phishing. Phishing remains one of the most prevalent forms of cybercrime. It exploits human trust to manipulate individuals into divulging sensitive information. This study investigates the trust development mechanisms most exploited by cybercriminals in phishing attacks. It focuses on two primary trust development processes: relationship history and future expectations. The study uses qualitative content analysis of 42 phishing messages collected from diverse secondary sources. The findings reveal that future expectations – such as promises of rewards, urgent requests, or threats of penalties – dominate phishing tactics. By contrast, relationship history mechanisms exploit the existing or fabricated relationships to evoke trust and compliance. These findings provide critical insights into the psychological manipulations leveraged in phishing schemes and highlight the need to integrate behavioural and cognitive principles into cybersecurity education. Practical implications include tailored training programs for distinct user groups, such as seniors, employees, and

Received: 12.09.2024

Accepted: 21.12.2024

Published: 30.12.2024

Cite this article as:

M. Daudi, "Exploiting human trust in cybersecurity: which trust development process is predominant in phishing attacks?," ACIG, vol. 3, no. 2, 2024, pp. 233–249. DOI: 10.60097/ACIG/199452

Corresponding author:

Morice Daudi, Computing Science Studies, Mzumbe University, Tanzania;
E-mail: dmorice@mzumbe.ac.tz

 0000-0001-7907-427X

Copyright:

Some rights reserved

(CC-BY):

Morice Daudi
Publisher NASK



students. The training should emphasise on recognising urgency cues, emotional manipulation, and verification strategies.

Keywords

phishing attacks, human trust, trust development processes, future expectations, cybersecurity

1. Introduction

Cybersecurity discourse has traditionally framed humans as a problem – susceptible to social engineering, prone to error, and easily manipulated. This framing, however, presents a limited view [1]. It is limited because the exploitation of humans as the weakest link in cybersecurity stems from the interplay of human psychology, social engineering tactics, and system usability. The theoretical challenge behind this problem focuses on how to mitigate the inherent vulnerabilities of human factors in the cyber landscape. Despite substantial investments in technological defences, human errors remain the leading cause of security breaches, contributing to as much as 90% of cybersecurity incidents [2, 3]. These errors arise from various sources, such as insufficient awareness, inadequate training, and susceptibility to psychological manipulation through social engineering tactics [2, 4]. Those human factors in cybersecurity are multifaceted and include intentional or unintentional actions that compromise security. For example, social engineering tactics exploit cognitive biases and psychological triggers, deceiving individuals into revealing confidential information or performing actions that undermine security protocols [2, 5]. These attacks leverage psychological principles like authority, reciprocity, and scarcity to manipulate victims [6, 7]. The susceptibility of individuals to such manipulation highlights the critical need for comprehensive cybersecurity education and the fostering of a security-aware culture within organisations [3, 4]. Therefore, combining technological solutions with insights into human behaviour is crucial for strengthening organisational resilience against emerging cyber threats [5, 8].

The literature provides varied perspectives on examining daily cybersecurity incidents involving phishing. Mitnick and Simon [9] discuss the manipulative tactics employed by cybercriminals and highlight the calculated exploitation of human emotions and cognitive biases. Hadnagy [10] explores how attackers exploit cognitive biases, trust, and social norms to manipulate individuals. Investigating the relationship between trust and cybersecurity

risks, Alhasan [4] reveals that higher trust increases risky cybersecurity behaviours across cultures. Additionally, Khan et al. [3] and Triplett [11] explore how human factors, including decision-making processes, organisational culture, and leadership contribute to insider threat. Despite these contributions, a gap remains in understanding the specific trust development process that cybercriminals rely on in phishing attacks. While the psychological and organisational dimensions of trust exploitation have been studied, there is limited focus on attackers' exact mechanisms and stages of trust development processes. This gap is critical, as understanding these processes could lead to more effective countermeasures. To this end, the present paper investigates the trust development processes most commonly employed by cybercriminals in phishing attacks. The study addresses the following research question: 'Which trust development process do cybercriminals most often exploit in phishing?' The study contributes to cybersecurity education by identifying the prevalent trust-building processes used in these exploits. This contribution empowers users to protect themselves better.

The remaining part of the paper is organised into seven sections. Section 2 discusses trust, phishing, and social engineering techniques for exploiting human trust. Section 3 outlines the methodology of this paper, followed by the presentation of results in Section 4. The findings presented in Section 5 are followed by their implications as discussed in Section 6. Section 7 provides practical recommendations. The paper ends with Section 8 by providing concluding remarks.

2. Literature Review

The present section comprises three subsections. It starts by discussing trust development processes (subsection 2.1), followed by the exploitation of human trust (subsection 2.2). Subsection 2.3 presents phishing techniques. The section ends by discussing social engineering techniques in subsection 2.4.

2.1. Trust Development Processes

Trust between parties evolves through specific processes. Before delving into these processes, it is essential to have a clear overview of the parties involved in trust transactions and the roles each party plays. For a trust exchange to be completed, two parties must be engaged: a trustor and a trustee. The trustor (e.g. a person) is an entity that develops a degree of reliance on another object

and accepts being vulnerable to the possible actions of that other object [12]. Similarly, the trustee (e.g. a person) is the party in whom the trust resides and can exploit the trustor's vulnerabilities [13]. To clarify further, the trustor is the party that puts its expectations in the other party, while the trustee is the party in which that expectation resides [14]. With this brief overview, the processes of trust development are discussed as follows.

Trust development processes can be understood through two primary mechanisms: relationship history and future expectations. Trust rooted in relationship history is built upon the experiences gained from past interactions between the trustor and the trustee [15]. 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, a reference from a third party is usually used to infer the development of trust [14]. Examples of bases of trust that employ relationship history include process-based, knowledge-based, and relational trust.

On the other hand, trust formed through future expectations is often driven by anticipated outcomes. Humans may trust the other party by relying on what they expect to gain after committing a trust transaction. This form of trust involves calculating the potential benefits and risks of engaging or not engaging in a particular trust transaction. Individuals assess whether entering a trusting relationship will yield favourable results or mitigate potential risks [14]. Examples of bases of trust that employ relationship history include calculus-based, deterrence-based, and competence-based trust [14, 15]. Both trust development processes (relationship history and future expectation) emphasise trust's dynamic nature.

2.2. Exploitations of Human Trust

Given the importance of trust in human interactions, cybercriminals exploit it as a key tactic in breaching cybersecurity. They leverage psychological principles, such as authority, reciprocity, and social proof to manipulate trust [1]. Those acts deceive individuals into compromising security systems. Trust exploitation is particularly effective because it taps into the inherent human tendency to trust familiar or authoritative sources [16].

One common method to exploit people's trust is phishing. Phishing relies heavily on manipulating human behaviour. In phishing attacks, cybercriminals craft messages that appear to originate

from trustworthy and legitimate sources to exploit the victim's inherent trust [4]. This deception is often amplified through urgent language or fabricated consequences. Through deception, individuals are compelled to respond quickly without fully verifying the communication's authenticity. The effectiveness of such attacks underscores the importance of raising awareness and educating individuals about the dangers of blindly trusting digital communications [7], particularly those that demand immediate action.

The cultural dimensions of trust also play a significant role in how individuals respond to phishing and other forms of deception. Research has shown that trust levels vary across cultures, with some cultures exhibiting higher baseline trust in digital communications [2]. Understanding these cultural differences is crucial for developing tailored cybersecurity strategies that address the specific trust-related vulnerabilities of different populations [6]. Generally, trust exploitation in cybersecurity highlights the relationship between psychology and technology. It also highlights the need for tactics combining technological protections with cultural and psychological knowledge.

2.3. Phishing Techniques in Cybersecurity

Phishing remains one of the most prevalent and effective techniques that cybercriminals employ to compromise cybersecurity. Phishing attacks typically involve sending fraudulent e-mails or messages that appear to come from legitimate sources [9]. The authors claim that those messages or e-mails lure individuals into providing sensitive information, such as passwords. The effectiveness of phishing lies in its ability to exploit basic human behaviour, such as trust and fear [10]. Trust and fear are sometimes triggered by falsified urgency and the authoritative nature of the messages. Despite widespread phishing awareness, the technique continues to evolve, becoming increasingly sophisticated and more challenging to detect [5].

Spear phishing, a more targeted form, has become a dangerous threat. Unlike typical phishing, which targets a large audience, spear phishing targets certain people or organisations [3]. It frequently relies on creating highly customised messages using data obtained from social media or other public sources. These messages are designed to appear credible and relevant to the recipient [11]. The precision and personalisation of spear phishing make it a formidable challenge for cybersecurity professionals, who must constantly adapt their defences to counter these evolving threats [4].

Phishing attacks have expanded beyond e-mail to include other communication platforms, such as SMS (smishing) and voice calls (vishing). These multi-vector attacks allow cybercriminals to simultaneously exploit different aspects of human behaviour and technological vulnerabilities [14]. For instance, smishing messages may appear to come from a trusted source, like a bank, and include a link that directs the victim to a fake website where their credentials are stolen [2]. Diversifying phishing techniques across multiple channels shows cybercriminals' adaptability and the need for comprehensive cybersecurity strategies covering many phishing vectors [7].

2.4. Social Engineering Techniques

Social engineering encompasses various tactics to manipulate individuals into divulging confidential information or performing actions compromising security. Social engineering is highly effective at breaching security systems because it exploits human psychological and cognitive biases [10]. Unlike traditional hacking, which targets technical vulnerabilities, social engineering exploits the human element, often seen as cybersecurity's weakest link [9]. By exploiting psychological principles like trust, authority, and reciprocity, social engineers can bypass technological defences and gain unauthorised access to systems or data [16].

Pretexting is a widely used social engineering technique where attackers create a fictitious scenario to deceive victims into revealing sensitive information. This often involves impersonating a trusted individual or authority figure, such as an IT support technician or a government official, to make the request appear legitimate [6]. The technique is particularly effective in organisational settings, where employees may feel obligated to comply with requests from perceived authorities [5]. The success of pretexting hinges on the attacker's ability to craft a convincing narrative that resonates with the victim's expectations and prior experiences [7].

Baiting is another common social engineering tactic that involves tempting victims with an enticing offer. The offer may comprise a gift to manipulate victims into actions compromising their security. This method exploits the human inclination for free or valuable items, often resulting in the spread of malware or the theft of sensitive information [2]. Baiting capitalises on individuals' curiosity and their tendency to take risks for potential rewards. Like other social engineering techniques, the success of baiting highlights the critical need for robust cybersecurity education that fosters skepticism and critical thinking in digital interactions [3].

3. Methodology

This study utilises a qualitative content analysis approach to investigate the trust development processes exploited by cybercriminals in phishing attacks. The research explores the two primary trust-building mechanisms – *relationship history* and *forthcoming expectations* – and their prevalence in phishing messages. The study categorises and analyses phishing messages to identify patterns and trends using these trust development processes. The data for this study was collected from various secondary sources, such as academic publications, cybersecurity reports, and online repositories of phishing messages. Specifically, phishing messages were extracted through search engines. These sources were chosen due to their comprehensive coverage of phishing tactics and their relevance to the research topic. Most spam messages from those sources are generic, which is considered a reference for many spam messages. A total of 42 phishing messages were selected for analysis to comprehensively represent various phishing tactics. These messages were intentionally chosen to capture the trust development processes related to relationship history and future expectations.

The phishing messages included in this study were purposely selected based on the relevance, variety, and recency criteria. Firstly, messages were included if they explicitly or implicitly involved trust development tactics to deceive the recipient. Secondly, a diverse set of messages was selected to cover different types of phishing attempts, such as those related to financial incentives, urgent requests, or personal relationships. Finally, preference was given to messages representative of contemporary phishing tactics to ensure that the findings are relevant to current cybersecurity challenges. The selected phishing messages were analysed using a thematic content analysis method. Each message was reviewed to identify the trust development process utilised – either relationship history or forthcoming expectations. Each phishing message was coded according to the identified trust development process. The frequency of each type was recorded and analysed to determine which process is more commonly exploited by cybercriminals.

To ensure the reliability of the analysis, two assistant researchers independently coded phishing messages. Any discrepancies in coding were discussed and resolved through consensus to mitigate potential biases in message classification. Validity was addressed by triangulating the findings with existing literature (such as that in Daud [14]) on phishing tactics and trust development processes in cybersecurity. The results were compared with previous studies

to ensure that the identified patterns align with established knowledge in the field. Furthermore, as the study utilised publicly available data from secondary sources, no personal information was collected or analysed. All sources of phishing messages were adequately cited, and care was taken to ensure that the analysis did not involve any unethical data manipulation.

4. Results

Table 1 presents the results of spam messages used mainly by cybercriminals. These messages were extracted from literature sources [14, 17–21]. Of the 42 spam messages, 33 were based on the future expectation trust development process. The remaining nine messages were based on the relationship history trust development process.

This study categorises spam messages presented in Table 1 into various groups: account verification, billing statements, credit card offers, customer service inquiries, family matters, job offers, and package delivery notifications (Figure 1). The most commonly identified categories were prizes or gift cards and account verification requests, each occurring for six times. This high frequency indicates that cybercriminals often focus on areas where individuals are likely to respond quickly, sometimes without exercising adequate caution. Family matters, package delivery, and internal revenue services ranked third, fourth, and fifth, respectively.

An in-depth analysis of spam messages indicates that cybercriminals frequently employ specific trust-building techniques to deceive their victims. Notably, 78.6% of the spam messages analysed were designed using the *future expectation* trust-building process. This process often promises future rewards or urgent actions, such as account verification or prize claims. It leverages urgency and anticipation to compel recipients to respond quickly without critically evaluating the legitimacy of the request. For instance, the following messages are classic examples of this approach:

‘Congratulations! You’ve won a \$500 Amazon gift card. Claim it here [Link]’

and

‘Your IRS tax refund is pending acceptance. Must accept within 24 hours: [Link]’

Table 1. Sample spam messages used by cybercriminals

No.	Spam messages	Trust development process
1.	Congratulations! You've won a \$500 Amazon gift card. Claim it here [Link].	Future expectation
2.	ACTION REQUIRED. Please verify your Bank of America account information to avoid a hold on your account. Click here to confirm: [Link].	Future expectation
3.	You've been overcharged for your 2021 taxes. Get your IRS tax refund here: [Link].	Future expectation
4.	Get delivery updates on your USPS order [Number] here: [Link].	Future expectation
5.	Thank you for paying last month's bill. We're rewarding our very best customers with a gift for their loyalty. Click here! [Link].	Future expectation
6.	Congratulations! Your credit score entitles you to a no-interest Visa credit card. Click here to claim: [Link].	Future expectation
7.	We've received your resume and would love to set up an online interview. Click here [Link] or call us at [Phone Number] at your earliest convenience.	Relationship history
8.	There's an issue with your payment information from your recent order [Order Number]. Take action now: [Link].	Future expectation
9.	We have detected suspicious activity on your Wells Fargo account. Log in at [Link] to update your account preferences and protect your information.	Future expectation
10.	Hi Grandpa, it's me – I've been in a car accident, and my parents aren't around. Can you please send me money so I can get home? You can wire funds to me here: [Link].	Relationship history
11.	'Your 2FA settings are not up to date. To avoid account suspension, please click the following link to update your settings: [Link]'.	Future expectation
12.	'Hey, it's [Boss Name]. I'm in a meeting now and need your help with something urgent. Can you transfer \$5,000 to this account ASAP? I'll explain everything later. Please keep this confidential'.	Relationship history
13.	'We're happy to inform you that you're entitled to a refund for overpayment on your AMEX account. Click on this link [Link] below to claim your refund'.	Future expectation
14.	Congratulations! You have all been selected to receive a free gift card worth \$1000. Click on this link [Link] to claim your reward now. Limited time offer, so act fast! Don't miss out on this amazing opportunity.	Future expectation
15.	'Congratulations! You've won a \$500 gift card to Target. Click here to claim your reward'.	Future expectation
16.	'Hello [Name], your shipment from UPS will arrive today. Click here to track your package'.	Future expectation
17.	'Your Wells Fargo account has been locked for suspicious activity. Please log in here and verify your account'.	Future expectation
18.	'Hey, this is [Name]. I'm in a meeting, but I need you to order 5 Amazon gift cards ASAP. I'll reimburse you once you send them to this e-mail address'.	Future expectation
19.	'[Name], your Verizon billing statement is ready. Please review your charges and send full payment by [date] to avoid late fees'.	Future expectation
20.	Congratulations! You've won a \$1000 Walmart gift card. Go to [Link] claim now.	Future expectation

(continues)

Table 1. Continued.

No.	Spam messages	Trust development process
21.	Your IRS tax refund is pending acceptance. Must accept within 24 hours: [Link].	Future expectation
22.	Wells Fargo Bank: Your account is temporarily locked. Please log in at [Link] to secure your account.	Future expectation
23.	Hello, your FEDEX package with tracking code DZ-8342-FY34 is waiting for you to set delivery preference: [Link].	Future expectation
24.	Apple Notification. Your Apple iCloud ID expires today. Log in to prevent deletion [Link].	Future expectation
25.	URGENT: Your grandson was arrested last night in Mexico. Need bail money immediately Western Union Wire \$9,500 [Link].	Relationship history
26.	Federal Credit Union ALERT: Your Credit Card has been temporarily LOCKED. Please call Card Services line [Tel. no].	Future expectation
27.	Thank you for your recent Amazon purchase. You've been charged \$108.34. If there has been a mistake, please call [Tel. no].	Future expectation
28.	Dear [Bank Name] customer, we've detected unusual activity on your account. Please click the link to verify your transactions: [malicious link].	Future expectation
29.	Hello, this is [Courier Service]. We've attempted to deliver your package today but failed. Schedule your redelivery here: [malicious link].	Future expectation
30.	We detected a login attempt from an unfamiliar location. If this wasn't you, please secure your account here: [malicious link].	Future expectation
31.	You're the lucky winner of our grand prize! Register here to receive your reward: [malicious link].	Future expectation
32.	A family member of yours has been in an accident. Call this premium rate number for details: [malicious phone number].	Relationship history
33.	I'm your landlord. My current number is unreachable. Send the rent through this number [Tel. no].	Relationship history
34.	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 expectation
35.	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 expectation
36.	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 expectation
37.	After unexpectedly collapsing at school, your son was brought to the hospital. Send money right away for medical care.	Relationship history
38.	Please get in touch with us as soon as you can; your child is extremely ill. Teacher.	Relationship history
39.	You won in the draw for the best customers who use our services. Please contact the following number to learn how to collect your prize.	Future expectation

(continues)

Table 1. Continued.

No.	Spam messages	Trust development process
40.	You have received Tshs 50,000 from [Tel. no] – (name of sender). New balance 67,850.00 Tshs. Trans ID: [Trans. No]. [Date and time].	Future expectation
41.	I’m at the funeral; please send twenty thousand shillings at the following phone number. I will pay back your money later.	Relationship history
42.	The person received a phone call from someone pretending to be a human resource officer at an airport. The caller claimed, the recipient’s job application had been received and requested Tsh 300,000 in exchange for persuading his superiors to select the recipient for the position.	Future expectation

Source: Extracted from [14, 17–21].

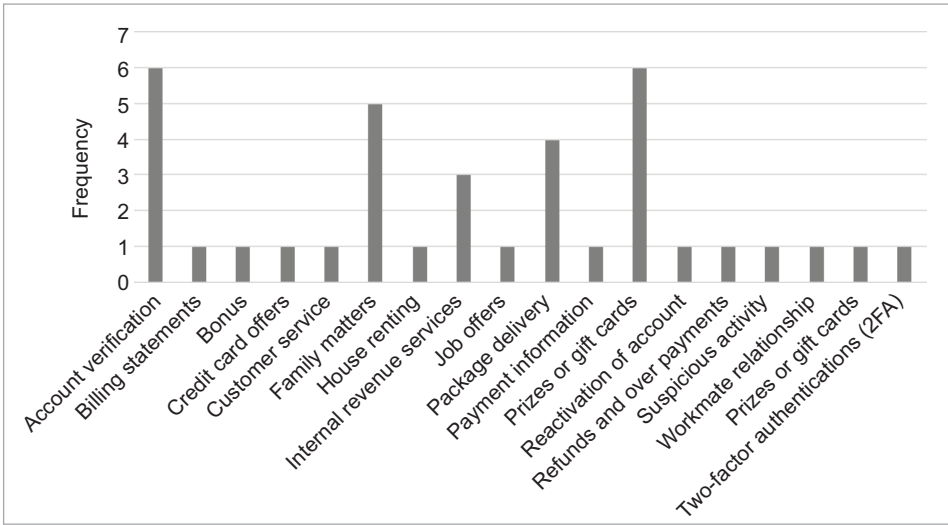


Figure 1. Categories of spam messages

These messages leverage the recipient’s hope for a positive outcome or fear of missing out.

On the other hand, the relationship history trust development process relies on exploiting the existing relationships or creating fictitious ones. These messages are crafted to appear as if they come from someone the recipient knows or trusts. They may appear to come from a family member, colleague, or service provider. An example of this would be the following message:

‘Hi Grandpa, it’s me – I’ve been in a car accident and my parents aren’t around. Can you please send me money so I can get home? You can wire funds to me here: [Link]’

The use of the relationship history trust development process emphasises the emotional connection and people's trust in their close relationships. It makes the recipient more likely to comply with the request without scepticism.

5. Discussion of the Findings

The analysis reveals key insights into cybercriminals' methods to exploit human trust in phishing attacks. One key insight concerns the prominence of the future expectation trust development process. Through future expectation, attackers target psychological triggers that urge immediate action. This approach is effective to attackers because it preys on common human behaviours. Examples of such behaviours include human tendency to seek financial gain or resolve issues quickly. Such behaviours are closely linked to the power of anticipation and urgency which cybercriminals understand and use it. Cybercriminals usually create scenarios where the victim believes they must act quickly to avoid negative consequences or secure a reward. While doing so, attackers limit the time available for critical assessment by the victim. This technique is particularly dangerous in today's fast-paced digital environment, where individuals often juggle multiple tasks and may overlook the need to scrutinise each message.

Various principles in literature underpin the dominance of future expectations as the trust-building process in phishing attacks. One is Cialdini's [16] principle of urgency and scarcity, where attackers create a sense of urgency, such as 'Your mobile money account will be closed immediately'. This tactic exploits the fear of missing out, pressuring victims to act quickly without assessing the message's legitimacy. This fear reinforces the deterrence-based trust developed through the future expectation process. Another principle is based on Sweller's [22] cognitive load theory, which posits that individuals under time pressure tend to rely on cognitive shortcuts (heuristics), rather than engaging in critical thinking. Attackers exploit this by leveraging the future expectation trust process. They do so by prompting victims to respond to phishing messages that promise rewards or threaten penalties. That action effectively bypasses the cognitive effort required to assess the authenticity of phishing messages.

Besides this, the findings from this study align with and extend to the existing research on phishing and social engineering in cybersecurity. For instance, an SMS phishing experiment revealed that combining urgency with either the promise of a reward or the

threat of a penalty successfully deceived 50% of participants [23]. This urgency is a key element in the future expectation trust development process discussed by Daudi [14]. Overall, these findings align with Vishwanath et al.'s [24] conclusion that superficial e-mail processing increases phishing success.

On the other hand, the use of relationship history as a trust-building process demonstrates the effectiveness of social engineering in phishing attacks. Cybercriminals bypass initial scepticism by impersonating someone the victim knows or trusts. This tactic exploits the victim's existing relationships, making it a powerful tool for attackers. It is particularly effective in urgent scenarios, such as requests for emergency funds to care for a sick child at school. The relationship history trust development process identified in this study further illustrates how cybercriminals exploit emotional connections to bypass rational scrutiny. To bypass rational scrutiny, cybercriminals often build rapport to gain trust and extract sensitive information. The success of such cybercriminals' attacks is backed up by humans' tendency to rely on familiar cues when assessing the authenticity of messages [25].

In addition to the trust-building process based on future expectations, gifts and financial incentives are often employed in the trust development process rooted in relationship history. For instance, spear-phishing e-mails that exploit a fabricated relationship history tend to achieve higher success rates than generic phishing e-mails [26]. Some of these e-mails create a sense of urgency by demanding immediate action from victims. Under such time pressure, individuals are more likely to overlook security protocols, skip essential steps, and make decisions that compromise cybersecurity [27]. Similarly, Razaq et al. [28] observed that fraudsters frequently pose as bank officials or government representatives, leveraging urgency to prompt swift compliance and establish trust with their victims.

6. Implications

The findings of this study have significant implications for cybersecurity practices, policy-making, and behavioural research. The dominance of future expectation mechanisms in phishing attacks highlights the need to address cognitive biases like urgency, anticipation, and risk perception in cybersecurity training. It must be recognised that individuals often fall victim to phishing because they are manipulated into prioritising immediate outcomes over critical evaluation. This insight necessitates integrating behavioural

and psychological theories, such as cognitive load theory [22] and temporal discounting [29], into awareness programs. This integration helps users better process suspicious messages. Moreover, the marginality of relationship history mechanisms suggests that attackers also capitalise on emotional connections. For this reason, users should exercise caution and verify communications, particularly those claiming personal relationships.

From a policy perspective, cybersecurity frameworks should incorporate behavioural training alongside technical solutions. Emphasis should be placed on vigilance and critical thinking. Additionally, the results indicate the need for adaptive cybersecurity measures that account for cultural and demographic differences in trust dynamics. Future research should explore these variations more deeply to develop region-specific strategies. Overall, this study emphasises that mitigating phishing effectively requires a holistic approach. This approach must integrate technological defences, psychological insights, and user education to create strong protection against evolving cyber threats.

7. Practical Recommendations

Organisations and individuals must implement targeted strategies to counter phishing attacks exploiting trust mechanisms. Firstly, cybersecurity training programs should focus on psychological manipulation tactics, such as urgency and anticipated rewards in phishing messages. These programs should teach individuals to recognise common phishing patterns, such as requests for immediate actions, financial rewards, or penalties. Secondly, organisations should simulate real-world phishing scenarios through controlled phishing campaigns. These exercises provide users with hands-on practice in identifying suspicious messages and offer immediate feedback. This approach effectively enhances their resilience against such attacks. Thirdly, automated e-mail and message filters should be strengthened by using appropriate tools. These tools can detect phishing-related language patterns, such as urgency cues or impersonation attempts. Verification practices should be emphasised for individuals. They should involve crosschecking of messages through alternative channels like direct calls or official websites. Lastly, organisations must develop user-specific awareness programs tailored to various demographics, such as employees, older adults, and students. This is because each group faces distinct vulnerabilities to trust-based phishing tactics. Combining these strategies will improve detection rates and minimise successful phishing exploits.

8. Conclusion

The exploitation of human trust to deceive and manipulate computer system users has become a significant concern in cybersecurity. Through social engineering and phishing, many users have fallen victim in various contexts. This study reveals that phishing attacks primarily exploit human psychological vulnerabilities through two trust development processes: future expectations and relationship history. The findings indicate that future expectations – such as promises of rewards, warnings of penalties, or urgent requests – are the most frequently used mechanisms by cybercriminals. These tactics rely on creating a sense of urgency and anticipation. Through this sense, victims are compelled to act impulsively without critically assessing the message's legitimacy. On the other hand, relationship history exploits familiarity and emotional connections. Attackers use this method to build trust by impersonating known individuals or organisations. The study highlights the need to incorporate behavioural insights into cybersecurity training and awareness programs. These programs should address cognitive biases, such as urgency and emotional triggers, to help individuals better identify and resist phishing attempts. Furthermore, mitigating phishing threats requires a multifaceted approach combining technological defences, user education, and understanding the human psychology of trust. By addressing these aspects holistically, individuals and organisations can develop more effective strategies to combat evolving phishing tactics and enhance overall cybersecurity resilience.

While this research provides valuable insights into the trust mechanisms exploited in phishing, its reliance on secondary data introduces certain limitations. Future studies should incorporate primary data collection methods, such as surveys, interviews, or experiments, to better understand user behaviours and responses to phishing attacks. Such approaches can provide richer insights into how cybercriminals exploit trust in real-world scenarios.

References

- [1] R. Anderson, *Security engineering: A guide to building dependable distributed systems*. Hoboken, NJ: Wiley, 2020.
- [2] A. Pollini et al., "Leveraging human factors in cybersecurity: An integrated methodological approach," *Cognition Technology & Work*, vol. 24, no. 2, pp. 371–390, 2022. doi: [10.1007/s10111-021-00683-y](https://doi.org/10.1007/s10111-021-00683-y).
- [3] N. Khan, R.J. Houghton, S. Sharples, "Understanding factors that influence unintentional insider threat: A framework to counteract unintentional risks,"

Cognition Technology & Work, vol. 24, no. 3, pp. 393–421, 2022. doi: [10.1007/s10111-021-00690-z](https://doi.org/10.1007/s10111-021-00690-z).

- [4] I. Alhasan, *Human factors in cybersecurity: A cross-cultural study on trust*. West Lafayette, IN: Purdue University, 2023.
- [5] S. Chaudhary, V. Gkioulos, S. Katsikas, “Developing metrics to assess the effectiveness of cybersecurity awareness program,” *Journal of Cybersecurity*, vol. 8, no. 1, pp. 1–19, 2022. doi: [10.1093/cybsec/tyac006](https://doi.org/10.1093/cybsec/tyac006).
- [6] E.O. Yeboah-boateng, P.M. Amanor, “Phishing, SMiShing & vishing: An assessment of threats against mobile devices,” *Journal of Emerging Trends in Computing and Information Sciences*, vol. 5, no. 4, pp. 297–307, 2014.
- [7] H. Kilavo, L.J. Mselle, R.I. Rais, S.I. Mrutu, “Reverse social engineering to counter social engineering in mobile money theft: A Tanzanian context,” *Journal of Applied Security Research*, vol. 18, no. 3, pp. 546–558, Jul. 2023. doi: [10.1080/19361610.2022.2031702](https://doi.org/10.1080/19361610.2022.2031702).
- [8] M. Grobler, R. Gaire, S. Nepal, “User, usage and usability: Redefining human centric cyber security,” *Frontiers in Big Data*, vol. 4, pp. 1–18, 2021. doi: [10.3389/fdata.2021.583723](https://doi.org/10.3389/fdata.2021.583723).
- [9] K. Mitnick, W.L. Simon, *The art of deception: Controlling the human element of security*. Hoboken, NJ: Wiley, 2002.
- [10] C. Hadnagy, *Social engineering: The art of human hacking*. Hoboken, NJ: Wiley, 2010.
- [11] W.J. Triplett, “Addressing human factors in cybersecurity leadership,” *Journal of Cybersecurity and Privacy*, vol. 2, no. 3, pp. 573–586, 2022. doi: [10.3390/jcp2030029](https://doi.org/10.3390/jcp2030029).
- [12] M. Daudi, *Trust in sharing resources in logistics collaboration*. Düren: Shaker Verlag GmbH, 2019.
- [13] 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](https://doi.org/10.1108/02621711211230858).
- [14] M. Daudi, “Trust framework on exploitation of humans as the weakest link in cybersecurity,” *Applied Cybersecurity & Internet Governance*, vol. 2, no. 1, pp. 1–26, 2023. doi: [10.60097/acig/162867](https://doi.org/10.60097/acig/162867).
- [15] N.P. Nguyen, N.T. Liem, “Inter-firm trust production: Theoretical perspectives,” *International Journal of Business, Management*, vol. 8, no. 7, 2013. doi: [0.5539/ijbm.v8n7p46](https://doi.org/0.5539/ijbm.v8n7p46).
- [16] R.B. Cialdini, *Influence: The psychology of persuasion*. NewYork, NY: Harper Business, 2007. doi: [10.1021/jf970693b](https://doi.org/10.1021/jf970693b).
- [17] J. Chantel. (2022). *10 spam text message examples (& how to identify them)* [Online]. Available: <https://blog.textedly.com/spam-text-message-examples> [Accessed: Mar. 06, 2024].
- [18] Proofpoint. (2023). *State of the phish report* [Online]. Available: <https://www.proofpoint.com/us/threat-reference/smishing> [Accessed: Apr. 17, 2024].

- [19] SlickText. (2023). 17 Spam text statistics & spam text examples for 2024 [Online]. Available: <https://www.slicktext.com/blog/2022/10/17-spam-text-statistics-for-2022/> [Accessed: Mar. 06, 2024].
- [20] R. Smith. (2023). "Stop scammers! 14 Examples of spam text messages," *Texting Base* [Online]. Available: <https://blog.textingbase.com/how-to-identify-spam-text-messages> [Accessed: Apr. 17, 2024].
- [21] I. H. Bakar (2016). *Social engineering tactics used in mobile money theft in Tanzania*, The University of Dodoma [Online]. Available: <http://repository.udom.ac.tz/handle/20.500.12661/1168> [Accessed: Jan. 09, 2024].
- [22] J. Sweller, "Cognitive load theory," *Cognition Science*, vol. 12, no. 2, pp. 257–285, 1988. doi: [10.1016/0364-0213\(88\)90023-9](https://doi.org/10.1016/0364-0213(88)90023-9).
- [23] H. Siadati, T. Nguyen, P. Gupta, M. Jakobsson, N. Memon, "Mind your SMSes: Mitigating social engineering in second factor authentication," *Computers & Security*, vol. 65, pp. 14–28, 2017. doi: [10.1016/j.cose.2016.09.009](https://doi.org/10.1016/j.cose.2016.09.009).
- [24] A. Vishwanath, T. Herath, R. Chen, J. Wang, H.R. Rao, "Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model," *Decision Support System*, vol. 51, no. 3, pp. 576–586, 2011. doi: [10.1016/j.dss.2011.03.002](https://doi.org/10.1016/j.dss.2011.03.002).
- [25] R. Dhamija, J.D. Tygar, M. Hearst, "Why phising works," in *Proceedings of CHI-2006: Conference on human factors in computing systems, April 2006*. New York, NY: ACM, 2006, pp. 581–590. doi: [10.1145/1124772.11248](https://doi.org/10.1145/1124772.11248).
- [26] K. Dubovecka, "Vulnerability of students of Masaryk University to two different types of phishing," *Applied Cybersecurity & Internet Governance*, vol. 4, no. 2, 2024. doi: [10.60097/ACIG/190268](https://doi.org/10.60097/ACIG/190268).
- [27] N.H. Chowdhury, M.T.P. Adam, T. Teubner, "Time pressure in human cybersecurity behavior: Theoretical framework and countermeasures," *Computers & Security*, vol. 97, p. 101931, Oct. 2020. doi: [10.1016/j.cose.2020.101931](https://doi.org/10.1016/j.cose.2020.101931).
- [28] L. Razaq, T. Ahmad, S. Ibtasam, U. Ramzan, S. Mare, "We even borrowed money from our neighbor," *Proceedings of ACM Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–30, Apr. 2021. doi: [10.1145/3449115](https://doi.org/10.1145/3449115).
- [29] R. Herrnstein, "Temporal discounting," *Journal of Experimental Analysis of Behavior*, vol. 4, no. 3, pp. 267–272, 1961. doi: [10.1901/jeab.1961.4-267](https://doi.org/10.1901/jeab.1961.4-267).

Denmark's Sector Responsibility Principle: A Tedious Cyber Resilience Strategy

Mikkel Storm Jensen | Royal Danish Defence College, Copenhagen, Denmark | ORCID: 0000-0003-3995-3020

Abstract

In 2014, Denmark launched its first national strategy for cyber resilience of critical infrastructure (CI). The 'National Cyber and Information Security Strategy' and its two subsequent successors from 2018 and 2022 follow the Sector Responsibility Principle (SRP). According to the principle, the state distributes the task of achieving and maintaining societal resilience to individual sectors, for example, health, energy supply, or finance, while maintaining central oversight and responsibility for implementation. Denmark is not alone in taking this approach: in fact, all the Nordic countries have applied some version of SRP. Danish governments have over the last decade taken significant steps to implement and facilitate societal cyber resilience through development of institutions, strategies, legal measures, and public-private partnerships (PPP). That said, Danish governments have gone less far than, for example, Finland's to take measures to achieve efficacy, and significant weaknesses are still left to be addressed. The article outlines the principles behind SRP and, using mainly Danish examples, demonstrates why implementation of SRP is both legally, organisationally, and technically difficult but also politically 'unpleasant'. Resilience is desirable but also a tedious chore. An inherent risk with SRP at both strategic, political level and individual private or public entity level are incentives to strive for legal compliance, rather than operational efficacy and act more according to a 'sector responsibility avoidance principle'. In that light, the article outlines how the SRP has been implemented in Denmark so far, along with examples

Received: 09.10.2023

Accepted: 10.04.2024

Published: 05.07.2024

Cite this article as:

M.S. Jensen, "Denmark's sector responsibility principle: A tedious cyber resilience strategy," ACIG, vol. 3, no. 2, 2024, pp. 250–267. DOI: 10.60097/ACIG/189870

Corresponding author:

Mikkel Storm Jensen, Royal Danish Defence College, Denmark; E-mail: msje@fak.dk

 0000-0003-3995-3020

Copyright:

Some rights reserved:
Publisher NASK



of both what drives the effort and challenges to successful SRP implementation.

Keywords

cyber, strategy, resilience, sector responsibility principle

1. Introduction

Danish governments have over the last decade taken significant steps to implement and facilitate societal cyber resilience through the development of institutions, strategies, legal measures, and public-private partnerships (PPP).¹ Denmark is not alone in taking this approach: in fact, all the Nordic countries have applied some version of Sector Responsibility Principle (SRP) [1]. In 2014, Denmark launched its first national strategy for achieving cyber resilience of critical infrastructure (CI). The ‘National Cyber and Information Security Strategy’ [2] and its two subsequent successors from 2018 [3] and 2021 [4] follow SRP. According to the principle, the state distributes the task of achieving and maintaining societal resilience to individual sectors, for example, health, energy supply, or finance, while maintaining central oversight and responsibility for implementation. However, Danish governments have gone less far than, for example, Finland’s to ensure the efficacy of the implemented strategies. According to North Atlantic Treaty Organization’s (NATO) 2020 evaluation, weaknesses in governance of resilience measures are still left to be addressed [5, p. 5]. This raises the question: why Denmark has not gone as far as Finland?

The literature on societal resilience strategies explains the sound principles behind SRP. This article seeks to add nuances to this body of literature by looking at the Danish case with an eye to identify incentives against implementing SRP with efficacy, rather than formal compliance as the main goal at both macro and micro levels.

After a literature review, the article outlines the principles behind SRP and demonstrates why it is a good strategic approach for states to achieve cyber resilience in modern, digitalised, and diverse economies. Methodologically, the article demonstrates why implementation of SRP in practice is not only legally, organisationally, and technically very difficult but also politically ‘unpleasant’ using mainly Danish examples. Denmark is a relevant case for studying potential weaknesses in cyber resilience strategies, as it is a highly digitalised society that has consistently scored high in international evaluations of national cybersecurity, although its position has

1——While the author is a serving officer with the Danish armed forces, the statements in this article are his own and do not present the position of the Danish Defence or the Danish Government.

fallen since ITU's initial evaluation in 2015 [6, 7]. The article takes its outset in the, so far, three Danish national information and cybersecurity strategies as well as the accompanying European Union (EU) NIS and NIS 2 directives. This presents methodological challenges: there are no formal definitions of a strategy, but according to, for instance Yarger and Bartholomees' [8] strategies should include political ends and explicit theories of success regarding assumed causalities between allocated means and appropriate ways. This allows observers to identify, assess, and discuss risks, for example, from potentially inadequate means or questionable ways and evaluate the theory of success' internal causality or compare with the result of other strategies in similar empirical contexts. Held to Yarger and Bartholomees' standards, the Danish strategies are lacking in content. Particularly the 2021 strategy [4] is mainly a list of aspirational ends, while ways and particularly allocated means are not specified in detail. This constitutes an analytical weakness, as the lack of explicit ways and means leaves a large amount to the external observer's interpretation. Even so, the approach gives indications as to where weaknesses may lay in the presented strategies, illustrated anecdotally with empirical observations from resilience-related events as they appear in reputable news sources or other reporting.

To governments as well as their citizens and enterprises, resilience is desirable but also a tedious chore that takes away resources from core services. An inherent risk with the SRP at both the strategic, political level and the individual private or public entity level is incentive to strive for legal compliance rather than operational efficacy and act more according to a 'sector responsibility avoidance principle'. Having discussed this in principle, the article will outline how the SRP has been implemented in Denmark so far, along with examples of both what drives the effort and challenges to successful SRP implementation.

2. Cyber resilience strategy – a new academic field

The article's headline includes the three concepts of 'strategy', 'cyber resilience', and 'sector responsibility principle', which the present literature goes some way to define. As mentioned above, Yarger and Bartholomees provide an operational definition of strategy as a formulated theory of success on how ends are achieved by applying sufficient means in particular ways. Furthermore, Yarger and Bartholomees provide a framework for describing the level at which strategies are developed and implemented. In the present

case, the investigated Danish strategies are at what Yarger and Bartholomees define as the 'National Security Strategy' level, as the means deployed include all aspects of the national instruments of power [8, pp. 48–49]. National cyber resilience strategies can encompass a number of relevant topics: building a cyber-workforce, promoting public cyber literacy, etc. This article focuses on the state's task of protecting critical infrastructure, particularly its role in developing and implementing strategy in the shape of institutions and regulations and how PPP is enforced, encouraged, and facilitated. Here, Tiirma-Klaar provides an overview of the areas that states may include in cyber resilience strategies [9]. Cyber resilience as such, particularly at the tactical level as the concept applies to individual entities and organisations, is described from many perspectives, and for instance, Sepúlveda Estay et al. provide oversight of relevant literature [10]. A search for 'sector responsibility principle' on Google Scholar, however, provides only Jensen [11] in spite of the principle's widespread use in Scandinavia [1].

Identifying the state's objective to be 'resilience' rather than 'security' is an acknowledgement of a governing principle, where the state is more a gardener guiding and facilitating a complex society's ability to withstand, overcome, and emerge stronger from external blows, than an engineer trying to keep external blows from affecting the societal machine or assist in repairing it afterwards. The emergence and history of this approach are well described by, for instance, Walker and Cooper [12]. This principle and the state's role therein is brilliantly described by Dunn-Cavelty and Suter in their article 'Public-private partnerships are no silver bullet: An expanded governance model for critical infrastructure protection' [13]. In this key piece, they describe how the strategic context for national resilience strategies has changed, particularly since the end of the Cold War. Modern economies used to be complicated, but some factors made it possible for the state to manage crisis through collection and analysis of information and central allocation of resources through commands, economic incentives, or patriotic encouragement [14, p. 2]. Critical infrastructure (CI) within, for example, production or communications was state-owned or run by domestic industries and based on standard communications systems like telephone, mail, order books, etc. This allowed a state to conduct 'business continuity management' (BCM) at a national level for extended periods. The world wars provided excellent examples of such state-run economies with 'PPP' based on central control [12, p. 3; 15]. But during the 1990s, many Western economies changed: state-run critical infrastructure was sold to private entities and these along with other domestic industries

often became international, either due to ownership or based on outsourcing from national or foreign subcontractors, always prone to change. At the same time, digitisation meant that command and control within critical infrastructure became based on innumerable and ever-changing systems [13, p. 180]. These and other changes transformed the basic structure of modern economies from complicated to complex, and made the hitherto successful central control approach to crisis management impractical [16, p. 46]. In the modern context the state's role is not to manage through direct intervention. The state's principal challenge is to create a framework that ensures – and facilitates – the individual sectors' resilience within critical infrastructure [13, pp. 183–186]. Only in the individual sectors are the necessary insights to identify, implement, and maintain resilience and overcome external blows [17]. Hence, the state must delegate the tasks involved to achieve resilience [18, p. 36; 19, p. 481]. Christensen and Lund-Petersen elaborate on the cyber aspects of PPP and resilience in 'Public-private partnerships on cyber security: A practice of loyalty' [20].

Dunn-Cavelty and Suter's analysis of meta-governance of self-organising networks identifies the state's tasks, thus: (1) define and communicating goals and priorities, (2) identify *status quo* and needs for action, (3) choose instruments, and (4) verify efficiency – and go to step 2 again [13, p. 185]. In practice, this means that to conduct meta-governance, a state should identify, designate, and keep track of CI, divided into sectors according to tasks to facilitate the emergence of networks. Also, it should set strategic objectives, for instance, through contracts, that sectors or individual suppliers must fulfil. Furthermore, set and enforce minimum standards, for example, ISO 27001 compliance, for cyber resilience in CI. And finally, it is important to facilitate PPP, for instance, by providing threat intelligence, promoting best practices, or improving access to reports and prosecuting cybercrime.

It is important to note that delegating the tasks does not mean delegating the responsibility: comprehensive security, including BCM of the nation's critical infrastructure, remains the state's responsibility towards its citizens even if the actual infrastructure involved has been sold to a private contractor [18, p. 37]. Furthermore, it is important to note that except for the financial sector, market forces are often insufficient to incentivise individual entities in CI, whether public or private, to achieve the levels of resilience that would be sufficient from a societal perspective [11, p. 5; 21, p. 266]. And, again it must be reiterated that the task of developing and implementing the necessary strategies is simple in principle, but

very difficult in practice and hampered by strong incentives that can lead to sub-optimisation at both strategic and individual levels. Dr. Kerttunen, who took part in developing Finland's comprehensive cyber resilience strategy, has expressed it thus:

What is the best strategy? It is relevant, optimized, updated, and implemented! There are three categories of states when it comes to cyber strategies: those without strategies, those with utopian strategies that cannot be implemented, and those with realistic strategies that are poorly implemented [1, p. 275; 22].

In Denmark, SRP is the guiding principle for resilience, including cyber resilience. This is stated by law and entails that the authority or institution, for instance ministry, who has the day-to-day responsibility for a task, also has the responsibility for planning, and resolving this task in a crisis [23, 24]. The fact that Denmark is now implementing its third cyber resilience strategy and has achieved some results, with its two predecessors placing Denmark in the third category of Dr. Kerttunen's conceptual framework. The next section elaborates on the strengths and weaknesses of the Danish approach.

3. Denmark's cyber resilience strategies

Since 2001, Denmark has had national strategies for the public sector's, citizens', and corporations' use of the cyber domain [25]. In 2014, the first national strategy for cyber and information security was introduced. It had set basic objectives, for instance, requiring ISO27001 implemented in government entities as well as some other concrete measures in identified CI in the telecommunications and energy sectors. Furthermore, it provided guidance to the newly established national Computer Emergency Response Team (CERT), Centre for Cyber Security (CFCS) under the Danish Defence Intelligence Service, and National Cyber Crime Centre (NC3) under the police, and initiated a program of information collection to establish *status quo* and identify major weaknesses [2]. The first strategy thus followed the model for meta-governance quite closely. The plan was to build on the results of this strategy with the introduction of a more extensive strategy in 2017. Developments were also driven forward by the introduction of the EU's Directive 2016/1148 concerning measures for a high common level of security of network and information systems – in daily terms, the NIS directive, which Denmark as an EU member was obliged to implement [26].

However, the initial plan did not hold. In 2016, the Ministry of Defence was tasked with developing a new strategy, and relevant ministries were ordered to participate in the process. However, after repeated delays, the government transferred the task to the Ministry of Finance. Likely, the lack of progress was due to the fact that efforts to develop individual ministries' contributions to the strategy had to compete with the ministries' core functions and were not given priority. In Denmark, the Ministry of Defence has no means to influence the quality and scale of other ministries' efforts. Also, while the Ministry of Defence was responsible for the cross-ministerial coordination, it was not provided extra funding with which to facilitate its progress. The Ministry of Finance has significantly more influence on other ministries through the power of the purse and a new strategy was eventually presented by an entity established under the ministry for the purpose, Digitaliseringsstyrelsen ('the Board for Digitization') in 2018 [3; 11, p. 10]. While Denmark has no official definitions of what constitutes CI, the commission for the strategy included designated sectors within which entities could be designated as CI, namely energy, health, transport, telecommunications, finance, and maritime transport. This was supplemented by the criteria for CI designation of the EU's NIS directive [3, pp. 38–40; 20, p. 3; 26, 27]. The 2018 strategy included both concrete initiatives to increase CI resilience but also initiatives to facilitate PPP. Part of the strategy was that each of the designated sectors should develop individual resilience strategies, a process that was completed by the end of 2018 [28]. Furthermore, the strategy introduced a central entity (a 'styregruppe' or 'control group') and an accompanying reporting framework with the task of staying informed on how the implementation progressed in individual sectors and facilitating the sharing of, for instance, best practice between sectors [3, pp. 43–45]. Like its predecessor, the 2018 strategy follows the recommendations of meta-governance by building on the information collected after the first strategy was implemented and focusing on concrete initiatives with stated deadlines to establish and facilitate the individual sector's ability to improve resilience, including PPP.

In December 2021, Digitaliseringsstyrelsen presented Denmark's current strategy [4]. Compared with its two predecessors, it is less concrete: more describing intents and ambitions than stating objectives and setting deadlines [21, p. 261]. The 2021 strategy outlines a continuation and expansion of the previous strategies, for example, by the establishment of decentralised cyber and information security entities (DCIS). It also expands the state's practical facilitation of individual citizen's and enterprise's cyber resilience, for example, by establishing a new hotline for identity

theft, strengthening the police's capability to prosecute cybercrime, and a special entity dealing with the cyber security challenges for small- and medium-sized enterprises (SMEs) that make up a significant part of the Danish economy [4, pp. 11, 14]. As such, the strategy continues to follow the principles of meta-governance, but its less concrete form and more aspirational formulations make it less immediately applicable. There is an underlying and accompanying set of documents that much more explicitly outlines the implementation of the strategy to the individual sector; however, while formally unclassified, these are not accessible to the public.

According to the strategy's preamble, the plan is to follow up with a new strategy in 2024. In this regard, it is interesting to observe what role Digitaliseringsstyrelsen, which has been leading the process since 2017, play. In December 2022, Digitaliseringsstyrelsen was removed from the Ministry of Finance's portfolio and formally made an independent ministry. However, a ministry is responsible for two diverse areas: digital governance and equal gender rights [29]. Recalling the Ministry of Defence's difficulties in moving the development of the second strategy forward in 2017, the new Ministry of Digital Governance and Gender Rights may experience similar challenges regarding a 2024 strategy.

4. Challenges to Denmark's implementation of SRP and cyber resilience

Recalling Dr. Kerttunen's quip about national cyber resilience strategies, at this point it is relevant to review what the principle challenges are to Denmark's implementation of its cyber resilience strategies through the SRP doctrine, and consider how they manifest themselves.

Initially, it must be fully acknowledged that developing, implementing, and maintaining national cyber resilience strategies is always going to be an extremely difficult task legally, economically, technically, organisationally, etc. Hence, the following sections are in no way intended as condescending vis-à-vis the attempts that are done. Furthermore, realising that the tasks involved are truly daunting, the analysis does not address these difficulties, but instead address the challenges presented by incentives for complacency at both political-strategic and individual level.

The nature of these challenges is perhaps best illustrated with an example from the United States: In May 2021, Colonial Pipeline, a private enterprise that delivers fuel to most of the US east coast,

was paralysed as a result of a ransomware attack conducted by Russian cybercriminals. As a result, fuel supplies immediately dropped by 45%. Seventeen states had to declare a state of emergency that in some areas lasted for weeks as transportation of persons and goods came to a halt. Forensics later assessed that the ransomware attack had been possible because Colonial Pipeline lacked basic cyber security measures in place [30–32]. What went wrong? Was the enterprise not designated as CI? Was there no resilience strategy in place? Was Colonial Pipeline not in compliance with regulations? It turned out that strategy was in place, and the enterprise was designated as CI complying its rules and regulations. However, those rules were basically that Colonial Pipeline should read the government's – here TSA's – recommendations, and then follow those if felt inclined to. Colonial had read the recommendations, and were thus in compliance. But it was not inclined not to follow them, hence they had no effect. The rules have now been changed [31, 33].

How could such an in hindsight obviously inefficient approach to cyber resilience be developed and implemented? There are four good reasons at play: (1) Designating CI is politically unpleasant; (2) requiring and upholding demands for CI is politically unpleasant; (3) having updated and detailed insight into CI's cyber resilience or lack thereof is politically unpleasant; and (4) paying for cyber resilience is generally unpleasant (for an extensive elaboration of these arguments, see Jensen [11, 34]). To go through these four drivers that incentivise neglect of resilience measures, cyber or otherwise, let us review them individually.

Designating CI is unpleasant: When the state designates a private or public entity as CI, it either implicitly or explicitly imposes some demands regarding resilience measures that non-CI entities are not subjected to. This imposes extra costs for the CI-designated entity that has to be covered either by adding to the price of the provided services or compensated in some manner. Hence, there is an economic incentive against designating infrastructure as CI that may counterbalance operational considerations.

In the Danish case, it may be difficult to demonstrate this challenge with regard to cyber resilience, but a look at Denmark's interpretation of EU's directive No. 2008/114/EF may illustrate how relevant decision makers may be reluctant to designate infrastructure as CI. The EU directive defines 'European critical infrastructure' or 'ECI' as 'critical infrastructure located in Member States the disruption or destruction of which would have a significant impact on at

least two Member States' [35, p. L345/77]. In Denmark's case, one could expect, for example, bridges across the straits, transnational power and internet cables, or Copenhagen Airport (CPH), the largest in Scandinavia, to be designated as ECI. However, as of 2022, no Danish infrastructure was ECI. Why? Because there are substitutes in principle if less so in reality: for instance if the bridge to Sweden breaks down, there is a ferry. From an operational perspective, this may make little sense and probably goes against the spirit behind the EU directive. However, this is how the ministries for transport and energy interpret the letter of the directive when they biannually report 'no ECI in Denmark' to Brussels. Thus, Denmark is in compliance with the directive but IT has no effect if the EU's intent IS to strengthen ECI's resilience [21, p. 263]. That said, compliance with the EU's NIS directive and the recently updated version, NIS 2 has been and will continue to be a very important driver of the implementation of cyber resilience measures in Denmark [26, 36]. In February 2024, the Danish Ministry of Defence stated that the implementation of NIS 2 in Denmark was delayed, but it is still expected to be in place in 2024 [37].

Requiring and upholding standards for CI is unpleasant: Not only do these demands add costs to the provided service as described above, but the demanding entity, here the government, also has to allocate resources to enforce and keep track of their implementation, a further draw on resources.

In this regard, the nature of the sector also plays a role. Within the governance sector, implementing resilience requirements should in principle be a question of issuing commands and expecting the entities to follow orders. However, in 2014, as part of Denmark's initial strategy, government agencies were ordered to implement the ISO 27001 standard by the end of 2016. Even so, by December 2022 only two-thirds had done so in spite of 'a high degree of attention from leaders on the task' [4, pp. 19–20; 38]. Hence, implementation of even relatively simple resilience requirements is not unproblematic even within the government and likely also not in other public sectors, for instance, health. In the financial sector, market forces drive cyber resilience and security in advance of governmental requirements. In the telecommunications and energy sector, the involved enterprises are private but highly concentrated to a few large entities that are very capable technically regarding cyber security and resilience which enables sparring on relevant requirements and their implementation between these entities and the government. The transport sector, on the other hand, is similarly

composed of private enterprises, but many are SMEs that often have little or no skills when it comes to cyber and their IT systems and potential vulnerabilities are very diverse.

Insight into status of resilience is unpleasant: Knowing that cyber resilience in CI is sub-par entails a political responsibility to react. Not knowing provides 'credible deniability' and the SRP can become 'a sector responsibility avoidance principle' if political leadership in case of incidents due to lack of resilience can get away with the excuse that according to SRP, it is the sector's and even individual entity's task to ensure sufficient resilience.

As mentioned, the Danish 2018 strategy put a framework in place for CI sectors to report to a central entity on progress on the implementation of resilience measures and share best practices [3, p. 45]. However, the framework does not set specific formats or timelines for reporting. Occasional interviews with entities involved in the process suggest that while such reporting takes place, it is with uneven intervals and in different formats across different CI sectors. The lack of central oversight and the accompanying lack of resilience measures enforcement in Denmark in even very critical CI were recently demonstrated in a highly critical report from 'Rigsrevisionen', the Danish Parliament's special investigations board. It states the following:

The cyber security resilience of the 13 critical IT systems selected for this study is not satisfactory. The resilience of one of the authorities, where Rigsrevisionen examined several IT systems, is particularly unsatisfactory. The consequence of inadequate cyber security resilience is that critical services provided by the public sector risk being either seriously disrupted or impossible to deliver. It should be noted that the level of cyber security resilience varies between the authorities in the study [39, p. 3].

This suggests that the Danish strategies do not go as far to gain insight into the status of cyber resilience as they could. For comparison, in Finland, the government has gone considerably further: they identified the problems presented by uneven reporting in 2015, and since 2017, Finnish CI sectors have reported monthly to the government's national security committee in a fixed format involving a 22-point matrix. This committee, established in 2013, conducts monthly meetings and submits an annual report to the president [40, 41].

Exacerbating the lack of central awareness, there is no overall authority tasked with coordinating the individual sector's planning and preparation between incidents [20, p. 1435]. Denmark's designated crisis management organisation only come together in extraordinary situations and only temporarily have the authority to deal with the effects of a crisis [23]. The tasks of coordinating individual sector's planning and preparation is delegated according to the SRP. But, as the example with implementation of ISO 27001 demonstrates, even under SRP, giving an order to implement resilience measures does not mean it is carried out – even within the public sector. With SRP's decentralised responsibility for the implementation of the upcoming cyber strategy follows that individual ministries must interpret what their responsibility entails [34]. At the same time, the ministries evaluate themselves when assessing whether their respective sectors live up to their interpretation of their responsibility. This introduces significant risk that the sectors do not have a shared understanding of their tasks and that they do not give them the same priority – a fact also noted above by Rigsrevisionen. Biannual national exercises since 2006 have consistently been highlighting this in their 'Conclusions' [42, p. 5; 43, p. 6].

Paying for resilience is unpleasant: Under most circumstances, cyber resilience is not the core business for neither public nor private entities. Hence, resilience measures take away human and capital resources from whatever that core business is. In public service sectors, for example, health, the societally optimal level of resilience is in no way influenced by market forces, and hence arbitrarily set by political leadership. In private sectors, market forces have some influence, but the economically optimal dedication of resources to resilience may be far less from an individual enterprise's perspective than from the general society if the failure of that enterprise results in significant costs, as second-order effects of its failure ripple through the economy. Consider, for example, a small de-icing company that is critical for the function of a major airport in winter. Their revenue, and hence market incentive to ensure BCM, comes nowhere near the cost to society if aircraft cannot take off on a winter day due to a cyberattack. Historically, only in the Danish financial sector, market forces have been sufficient to drive cyber resilience to a very high level [44]. In the case of public sector, the political level can decide how much resources are taken from other tasks and dedicated to resilience, but who and how should the difference between the general society's and the small airport enterprise's incentive to invest in resilience be covered?

Recent research indicates that cyber security and resilience are often not a high priority in Denmark's many SMEs. In some cases, this is because implementation appears economically and/or technically challenging. In other cases, the task is too far from the experience and expertise of SMEs' leadership to rise to a sufficient level of attention to result in taking action [45, 46]. Since the introduction of the first national Danish strategy in 2014, Danish governments have primarily placed funding for implementation on the defence budget [47, p. 13; 48, p. 11]. This is in light of the magnitude of the task likely insufficient to cover the actual costs in all sectors. For instance, the Confederation of Danish Industry (Dansk Industri, DI) that promotes the interests of the SME sector assessed it as unlikely that the allocated 270 mio. DKK were sufficient to cover the 34 initiatives presented in the 2021 strategy [4, p. 5; 49].

5. The SRP is the right principle for Danish cyber resilience, but demonstrated political priority does not fully match stated ambitions

As the examples of this article have demonstrated, the state's role in establishing and maintaining comprehensive cyber resilience in CI is both highly complex and fraught with political and economic incentives to give the task less priority than a purely operational perspective might recommend. The Russian full-scale invasion of Ukraine in February 2022 has accentuated the need for resilience and the state's role in that regard. Denmark's national CERT has, along with other Western intelligence services, warned about an increased risk of Russian 'hactivism', and Danish banks, airports, ministries, and other CI have been the target for Russian distributed denial-of-service (DDOS) attacks [50–55].

The Danish strategies have, since 2014, along with EU's NIS directives, established a framework for solving the task. The strategies have, like in the rest of Scandinavia, built on SRP and contain the elements necessary to replace the state's role as 'the societal engineer' of the past with 'the societal gardener' of today and tomorrow. Governments from both sides of the parliament have built on their predecessors' strategies to establish institutions and frameworks to, for instance, identify and designate CI, assess the level of resilience, provide threat warning, and facilitate PPP. Also, the latest strategy's focus on SME opened a new and important area for implementing measures for cyber resilience.

However, as demonstrated by the examples, the implemented policies have still been insufficient to overcome incentives to give the

task less than the necessary priority, even within the public sectors, as demonstrated by the limited progress of ISO27001 implementation and the serious deficiencies in CI systems identified by Rigsrevisionen. SRP is the proper tool for the task, but the inherent threat from implementing it as the ‘sector responsibility avoidance principle’ has yet to be overcome – a challenge that Denmark shares with all Nordic countries that apply SRP [1, p. 274]. Ambitious headlines in the current and coming strategies do not decide the outcome. Only the government’s will and tenacity actually implement resilience measures through oversight, control, facilitation, guidance, and resource allocation.

References

- [1] M.S. Jensen, “Cyberresiliens, sektorprincip og ansvarsplacering – nordiske erfaringer,” *Internasjonal Politikk*, vol. 77, no. 3, pp. 266–277, 2019, doi: [10.23865/INTPOL.V77.1369](https://doi.org/10.23865/INTPOL.V77.1369).
- [2] Regeringen. (Dec. 2014). *National strategi for cyber-og informationssikkerhed – Øget professionalisering og mere viden*. København. [Online]. Available: <http://www.fmn.dk/nyheder/Documents/National-strategi-for-cyber-og-informationssikkerhed.pdf>. [Accessed: Aug. 19, 2020].
- [3] Finansministeriet. (2018). *National strategi for cyber-og informationssikkerhed*, Finansministeriet. [Online]. Available: <http://www.fmn.dk/nyheder/Documents/National-strategi-for-cyber-og-informationssikkerhed-2018.pdf>. [Accessed: Aug. 19, 2020].
- [4] Regeringen. (Dec. 2021). *The Danish National Strategy for Cyber and Information Security*, Regeringen. [Online]. Available: https://www.cfcs.dk/globalassets/cfcs/dokumenter/2022/ncis_2022-2024_en.pdf. [Accessed: Jan. 8, 2023].
- [5] North Atlantic Treaty Organization (NATO). (2020). *NATO Defence Planning Capability Review 2019/2020 Denmark C-M (2020) 0026 (DK-Overview)*, NATO. [Online]. Available: <https://www.fmn.dk/globalassets/fmn/dokumenter/aarsrapporter/nato/-na-to-defence-planning-capability-review-2019-2020-.pdf>. [Accessed: Feb. 6, 2023].
- [6] UN International Telecommunication Union (ITU). (2015). *Global Cybersecurity Index 2014*, ITU, Geneva. [Online]. Available: https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-SECU-2015-PDF-E.pdf. [Accessed: Feb. 29, 2024].
- [7] UN International Telecommunication Union (ITU). (2021). *Global Cybersecurity Index 2020*, ITU, Geneva. [Online]. Available: https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2021-PDF-E.pdf. [Accessed: Feb. 29, 2024].
- [8] H.R. Yarger, J.B. Bartholomees, “Toward a Theory of Strategy: Art Lykke and the U.S. Army War College Strategy Model,” *Strategic Studies Institute, US Army War College*, 2012. [Online]. Available: <https://www.jstor.org/stable/resrep12116.6>. [Accessed: Jan. 13, 2021].
- [9] 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](https://doi.org/10.1080/23738871.2016.1165716).

- [10] D.A. Sepúlveda Estay, R. Sahay, M. B. Barfod, C.D. Jensen, "A systematic review of cyber-resilience assessment frameworks," *Computers & Security*, vol. 97, p. 101996, 2020, doi: [10.1016/j.cose.2020.101996](https://doi.org/10.1016/j.cose.2020.101996).
- [11] M.S. Jensen, "Sector responsibility or sector task? New cyber strategy occasion for rethinking the Danish sector responsibility principle," *Scandinavian Journal of Military Studies*, vol. 1, no. 1, pp. 1–18, 2018, doi: [10.31374/sjms.3](https://doi.org/10.31374/sjms.3).
- [12] J. Walker, M. Cooper, "Genealogies of resilience: From systems ecology to the political economy of crisis adaptation," *Security Dialogue*, vol. 42, no. 2, pp. 143–160, 2011.
- [13] M. Dunn-Cavelty, M. Suter, "Public-private partnerships are no silver bullet: An expanded governance model for critical infrastructure protection," *International Journal of Critical Infrastructure Protection*, vol. 2, no. 4, pp. 179–187, 2009, doi: [10.1016/j.ijcip.2009.08.006](https://doi.org/10.1016/j.ijcip.2009.08.006).
- [14] K.C. Lauts, R. Hoffmann, L.B. Struwe, "Cyberwarfares Udfordringer af Begrebet Kritisk Infrastruktur," Københavns Universitet, Center for Militære Studier, Copenhagen, 2013. [Online]. Available: <http://curis.ku.dk/ws/files/66128849/Cyberwarfare.pdf>. [Accessed: Aug. 19, 2020].
- [15] S. Broadberry, M. Harrison, "The economics of World War I: An overview," in *The Economics of World War I*, M. Harrison and S. Broadberry, Eds., Cambridge University Press, Cambridge, 2005, pp. 3–40, doi: 10.1017/CBO9780511497339.002.
- [16] M. Carr, "Public-private partnerships in national cyber-security strategies," *International Affairs*, vol. 92, no. 1, pp. 43–62, 2016, doi: [10.1111/1468-2346.12504](https://doi.org/10.1111/1468-2346.12504).
- [17] J. Brassett, N. Vaughan-Williams, "The politics of resilience from a practitioner's perspective: An interview with Helen Braithwaite OBE," *Politics*, vol. 33, no. 4, pp. 229–239, 2013, doi: [10.1111/1467-9256.12027](https://doi.org/10.1111/1467-9256.12027).
- [18] J. Brassett, N. Vaughan-Williams, "Security and the performative politics of resilience: Critical infrastructure protection and humanitarian emergency preparedness," *Security Dialogue*, vol. 46, no. 1, pp. 32–50, 2015, doi: [10.1177/0967010614555943](https://doi.org/10.1177/0967010614555943).
- [19] M. Duffield, "Challenging environments: Danger, resilience and the aid industry," *Security Dialogue*, vol. 43, no. 5, pp. 475–492, 2012, doi: [10.1177/0967010612457975](https://doi.org/10.1177/0967010612457975).
- [20] K.K. Christensen, K.L. Petersen, "Public-private partnerships on cyber security: A practice of loyalty," *International Affairs*, vol. 93, no. 6, pp. 1435–1452, 2017, doi: [10.1093/ia/iix189](https://doi.org/10.1093/ia/iix189).
- [21] M.S. Jensen, "Cyberresiliens og kritisk infrastruktur: Vanskelige udfordringer og trøste løsninger," in *Cybertrusler: Det Digitale Samfunds Skyggeside*, Jeppe T. Jacobsen and Tobias Liebetrau, Eds., Djøf Forlag (Jurist og Økonomforbundets Forlag), Copenhagen, 2022.
- [22] M. Kerttunen, Lecture at George C. Marshall European Center for Security Studies, Garmisch-Partenkirchen, Germany, Dec. 06, 2021.
- [23] Beredskabsstyrelsen (BRS). (2022). *Krisestyringssystemet i Danmark*, Beredskabsstyrelsen. [Online]. Available: <https://www.brs.dk/da/arbejdsopgaver/om-krisestyring-og-redningsberedskabet/krisestyringssystemet-i-danmark/>. [Accessed: Apr. 22, 2022].

- [24] Forsvarsministeriet. (2017). Bekendtgørelse af beredskabsloven, vol. LBK nr 314 af 03/04/2017. Forsvarsministeriet. [Online]. Available: <https://www.retsinformation.dk/eli/lt/2017/314>. [Accessed: Sep. 29, 2021].
- [25] Digitaliseringsstyrelsen. (2023). *The Danish Digital Journey*, Digitaliseringsstyrelsen. [Online]. Available: <https://en.digst.dk/policy/the-danish-digital-journey/>. [Accessed: Aug. 17, 2023].
- [26] European Union (EU). (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*, European Parliament. [Online]. Available: <http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016L1148&from=EN>. [Accessed: May 13, 2018].
- [27] Forsvarsministeriet, “Kommisorium for the tværministerielle arbejde med den nationale strategi for cyber-og informationssikkerhed 2017–2019,” Danish Ministry of Defence, København (Copenhagen), 2016.
- [28] Forsvarsministeriet. (2019). *Nye sektorstrategier skal ruste samfundet mod cyberangreb*, Forsvarsministeriet. [Online]. Available: <https://fmn.dk/da/nyheder/2019/2019/nye-sektorstrategier-skal-ruste-samfundet-mod-cyberangreb/>. [Accessed: Sep. 30, 2021].
- [29] Ministry of Digital Governance and Gender Rights. (2023). *Ministry of digital governance and gender rights*, DIGMN. [Online]. Available: <https://english.digmin.dk/>. [Accessed: Aug. 17, 2023].
- [30] D.E. Sanger, N. Perlroth, “Colonial pipeline hack reveals weaknesses in US cybersecurity,” *The New York Times*, 2021. [Online]. Available: <https://www.nytimes.com/2021/05/14/us/politics/pipeline-hack.html?referringSource=articleShare>. [Accessed: May 17, 2021].
- [31] W. Turton, K. Mehrota, “Hackers breached colonial pipeline using compromised password,” Bloomberg.com, Jun. 04, 2021. [Online]. Available: <https://www.bloomberg.com/news/articles/2021-06-04/hackers-breached-colonial-pipeline-using-compromised-password>. [Accessed: Sep. 29, 2021].
- [32] D. Uberti, “TSA pipeline oversight faces scrutiny after colonial hack,” *Wall Street Journal*, May 13, 2021. [Online]. Available: <https://www.wsj.com/articles/tsa-pipeline-oversight-faces-scrutiny-after-colonial-hack-11620898202>. [Accessed: Sep. 29, 2021].
- [33] Republican Policy Committee (RPC). (2021). *Infrastructure cybersecurity pipelines*, RPC. [Online]. Available: <https://www.rpc.senate.gov/policy-papers/infrastructure-cybersecurity-pipelines>. [Accessed: Jan. 6, 2022].
- [34] M.S. Jensen. (Nov. 1, 2017). Author’s interview with Center for Cyber Security.
- [35] European Union (EU). (2008). *Raadets Direktiv 2008/114/EF af 8. december 2008 om indkredsning og udpegning af europæisk kritisk infrastruktur og vurdering af behovet for at beskytte den bedre*, EU. [Online]. Available: <http://eur-lex.europa.eu/legal-content/DA/TXT/PDF/?uri=CELEX:32008L0114&from=DA>. [Accessed: Mar. 9, 2017].
- [36] European Union (EU). (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 (NIS 2 Directive)*,

- European Parliament. [Online]. Available: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022L2555&qid=1692616271604>. [Accessed: Aug. 21, 2023].
- [37] Center for Cyber Security (CFCS). (2024). *Nye regler om cybersikkerhed bliver forsinkede*, Forsvarsministeriet. [Online]. Available: <https://www.fmn.dk/da/nyheder/2024/nye-regler-om-cybersikkerhed-bliver-forsinkede/>. [Accessed: Mar. 21, 2024].
- [38] Digitaliseringsstyrelsen. (2017). *Resultatet af undersøgelse af status paa implementering af ISO27001-principper i staten*, Digitaliseringsstyrelsen, København. [Online]. Available: <https://digst.dk/media/16012/resultat-for-staten-2017.pdf>. [Accessed: Apr. 9, 2018].
- [39] Rigsrevisionen. (2022). *Report on the cyber security resilience of the public sector*, Folketinget, Rigsrevisionen. [Online]. Available: <https://uk.rigsrevisionen.dk/audits-reports-archive/2022/nov/report-on-the-cyber-security-resilience-of-the-public-sector>. [Accessed: Aug. 21, 2023].
- [40] Turvallisuuskomitea. (2017). *Implementation programme for Finland's cyber security strategy*, Security Committee Finland Ministry of Defence, Helsinki. [Online]. Available: <https://www.turvallisuuskomitea.fi/index.php/en/component/k2/132-implementation-programme-for-finland-s-cyber-security-strategy-for-2017-2020>. [Accessed: Sep. 22, 2018].
- [41] Finland Security Committee. (2015). *Secure Finland – Information on comprehensive security in Finland*, Helsinki, Finland Security Committee. [Online]. Available: <https://www.turvallisuuskomitea.fi/index.php/en/component/k2/47-secure-finland-information-on-comprehensive-security-in-finland>. [Accessed: Apr. 4, 2017].
- [42] Beredskabsstyrelsen (BRS). (2017). *Evaluering af krisoer 2017*, Beredskabsstyrelsen. [Online]. Available: <https://www.brs.dk/globalassets/brs---beredskabsstyrelsen/dokumenter/krisestyling-og-beredskabsplanlagning/2020/evaluering-af-krisoer-2017.pdf>. [Accessed: May 26, 2022].
- [43] Beredskabsstyrelsen (BRS). (2016). *Evaluering af KRISOV 2015*, BRS. [Online]. Available: <https://www.brs.dk/globalassets/brs---beredskabsstyrelsen/dokumenter/krisestyling-og-beredskabsplanlagning/2020/evaluering-af-krisoer-2015-.pdf>. [Accessed: Jan. 06, 2022].
- [44] S. Goll. (2022). *Nordisk samarbejde i finanssektoren styrker kampen mod cyberkriminalitet*, Finans Danmark. [Online]. Available: <https://finansdanmark.dk/nyheder/2017/nordisk-samarbejde-i-finanssektoren-styrker-kampen-mod-cyberkriminalitet/>. [Accessed: Jan. 6, 2022].
- [45] V. Arildsen. (2021). *Ny rapport: Cybersikkerhed er underprioriteret i danske virksomheder*, UN International Telecommunication Union (ITU). [Online]. Available: <http://itu.dk/Om-ITU/Presse/Nyheder/2021/Ny-rapport-cybersikkerhed-er-underprioriteret-i-danske-virksomheder>. [Accessed: Jan. 06, 2022].
- [46] O. Kulyk, J. Mauro. (Dec. 2020). *Assessment on the status of cybersecurity in Denmark*, SDU, Odense, Dec. [Online]. Available: <https://ascd.dk/results/report.pdf>. [Accessed: Jan. 06, 2022].
- [47] T. Bramsen. (2021). *Et styrket dansk cyberforsvar*, Forsvarsministeriet. [Online]. Available: <https://fmn.dk/globalassets/fmn/dokumenter/nyheder/2021/-et-styrket-dansk-cyberforsvar-2021-ua-.pdf>. [Accessed: Jul. 15, 2021].

- [48] Regeringen. (Jan. 28, 2018). *Forsvarsforlig 2018*, Regeringen. [Online]. Available: <https://www.regeringen.dk/nyheder/2018/forsvarsforlig-2018/>. [Accessed: Jan. 29, 2021].
- [49] J.Ø. Wittorff. (2021). *Erhvervsorganisationer alvorligt bekymret over Danmarks nye cyber-strategi: Det er ikke nok*, Computerworld. [Online]. Available: <https://www.computerworld.dk/art/259045/erhvervsorganisationer-alvorligt-bekymret-over-danmarks-nye-cyber-strategi-det-er-ikke-nok>. [Accessed: Jan. 06, 2022].
- [50] L. Friis. (2022). *Københavns Lufthavn advarer efter flere russiske hackerangreb: Vi har en galoperende cyberrisiko*, Berlingske.dk. [Online]. Available: <https://www.berlingske.dk/content/item/1688115>. [Accessed: Aug. 21, 2023].
- [51] E.K. Stephensen. (2023). *Forsvarsministeriet udsat for cyberangreb – TV 2*, nyheder.tv2.dk. [Online]. Available: <https://nyheder.tv2.dk/samfund/2023-05-12-forsvarsministeriet-udsat-for-cyberangreb>. [Accessed: Aug. 21, 2023].
- [52] I. Meesenburg. (2023). *Finansministeriet ramt af cyberangreb*, DR. [Online]. Available: <https://www.dr.dk/nyheder/seneste/finansministeriet-ramt-af-cyberangreb>. [Accessed: Aug. 21, 2023].
- [53] M. Mezouri. (2023). *Prorussisk hackergruppe står bag angreb mod danske banker, siger it-sikkerhedseksperter og Danske Bank – TV 2*, nyheder.tv2.dk. [Online]. Available: <https://nyheder.tv2.dk/tech/2023-01-10-prorussisk-hackergruppe-staar-bag-angreb-mod-danske-banker-siger-it>. [Accessed: Aug. 21, 2023].
- [54] Center for Cybersikkerhed (CFCS). (2022). *Cybertruslen mod Danmark i lyset af Ruslands invasion af Ukraine*, CFCS. [Online]. Available: <https://www.cfcs.dk/da/cybertruslen/trusselsvurderinger/ruslands-invasion-af-ukraine/>. [Accessed: Apr. 19, 2022].
- [55] Cybersecurity & Infrastructure Security Agency (CISA). (2022). *Russian state-sponsored and criminal cyber threats to critical infrastructure*, CISA.gov. [Online]. Available: <https://www.cisa.gov/uscert/ncas/alerts/aa22-110a>. [Accessed: Jun. 18, 2022].

Vulnerability of Students of Masaryk University to Two Different Types of Phishing

Klara Dubovecka | Department of Political Science, Faculty of Social Studies,
Masaryk University, Czech Republic | ORCID: 0009-0003-9679-5767

Abstract

According to the European Union Agency for Cybersecurity's (ENISA) Threat Landscape (ETL) report 2020, phishing is the most commonly used type of cyberattack. Phishing is the technique of delivering false communications that appear to be from a real and respectable source, typically via e-mail or text message. The attacker aims to steal money, obtain access to sensitive data, and login information, or install malware on the victim's device. Data from the same report shows that during the COVID-19 pandemic, phishing attacks increased by 667% in one month. Simultaneously, warnings about expected waves of phishing e-mails at Masaryk University in Czechia were encountered more often. However, at the time this article was written, there was *de facto* no anti-phishing research dealing with the problem of phishing attacks on Czech universities. The present article focuses on unintentional human error on the side of students of Masaryk University. The main aim of this article is to uncover the profile of the user who is most prone to victimisation of phishing in the university setting. These results were achieved by performing two real-life phishing simulations. Data suggests that female students are more prone to crash for targeted e-mails. At the same time, all students are more susceptible to spear-phishing attacks than to the generic ones. Findings are explained by analysing the empirical results of the two real-life phishing attacks conducted.

Received: 4.11.2023

Accepted: 20.06.2024

Published: 24.07.2024

Cite this article as:

K. Dubovecka
"Vulnerability of students of Masaryk University to two different types of phishing," ACIG, vol. 3, no. 2, pp. 268–285. 2024, DOI: 10.60097/ACIG/190268

Corresponding author:

Klara Dubovecka,
Department of Political Science, Faculty of Social Studies, Masaryk University, Czech Republic; E-mail: dubovecka.klara@gmail.com;

 0009-0003-9679-5767

Copyright:

Some rights reserved:
Publisher NASK



Keywords

phishing, university students, social engineering, the human factor, unintentional threat

1. Introduction

We live in the information age, where connected devices and end-users increase daily [1]. The number has significantly risen with the COVID-19 pandemic because of home offices, online education, and entertainment via platforms during leisure time [2]. However, cybersecurity education rarely preceded this shift, which exposed a big group of end-users to cyberattacks daily. Different devices and technologies are used in people's personal lives, the companies they work for, the universities they study at, and the political institutions that govern them. Nevertheless, institutions cannot solely rely on a technological aspect of cybersecurity because of its interdependence. The importance of the human factor is still present, and the threat is growing simultaneously with the number of institutions that undergo speeded digital transformation, which is more than ordinary during these strange times of the COVID-19 pandemic. Institutions try to maintain the quality of virtual communication and, simultaneously, assure security in cyberspace while the shift has increased remote activities on the Internet. Human error continues to be the weakest link of cybersecurity – intentionally or unintentionally [3]. This vulnerability creates many opportunities for cybercriminals to attack human perception, rather than security measures through social engineering. Social engineering techniques trick individuals or organisations into accomplishing actions that benefit attackers or provide them with sensitive data [4]. The 2021 Data Breach Investigations Report (DBIR) states that social engineering has been responsible for 37% of all breaches in 2021 [5].

Notably successful were phishing attacks. A phishing attack is a cyberattack that exploits human vulnerability by disguising oneself as a trustworthy entity to influence or gain private information by sending an e-mail [6]. According to social engineers, 90% of all sent e-mails (294 billion each day) are spam and viruses, which means that e-mail is a significant vulnerability. The Anti-Phishing Working Group reported that phishing attacks hit an all-time high in December 2021 (316,747 attacks per month), meaning that phishing attacks have tripled since the early 2020 [7, p. 2]. Data from last year shows us that phishing aimed at the education sector is

increasing [5]. Universities are often a target, mainly because they store private and financial information and academic records of thousands of students and members of faculties. One disadvantage is their transient nature, which makes education about cybersecurity malpractices more complicated [8]. The vulnerability posed by phishing is often used effectively to the largest extent. Therefore, universities have a significant interest in protecting themselves from malicious cyberattacks. At a time when phishing is still in the limelight and the success rate of attacks on universities is increasing, there is virtually no research in the academic environment of the Czech Republic that focuses on the vulnerability of students and their ability to guess whether an email is legitimate or fraudulent.

This article focuses on unintentional human error and its threat to institutional cybersecurity by conducting real-life phishing simulations. The research goal of this experiment is to assess the profile of a student who is most susceptible to phishing and to provide the foundation for understanding how vulnerable students of the Faculty of Social Studies (FSS) at Masaryk University are. Phishing vulnerability is compared in two categories of e-mails – one generic and another targeted (spear-phishing). The research is limited to the Faculty of Social Studies at Masaryk University in Brno due to limited resources for this research. At the same time, this research provides a basis for similar research on a larger scale in the future.

The remainder of this article is organised as follows. Section 2 provides a literature review summarizing previous studies; Section 3 describes the methodology employed; Section 4 presents the results and analysis of data; and the final section discusses the implications of the findings.

2. Literature Review

The first phishing e-mail was sent in 1990 [9]. Fast forward 20 years, and it is the most commonly used tool for compromising an institution [10]. Many information security scholars have found phishing in a university community a research area of interest. Although studies have been performed before also, the most significant momentum has occurred in the past decade. Researchers have begun exploring what could be the user profile of a person most likely to react to phishing. Because of this, studies to capture demographics connected to phishing susceptibility have been administered in different universities worldwide. In the following section, related studies are discussed.

2.1. Phishing in General

Jerry Felix and Chris Hauck first outlined phishing as a strategy in which a third party imitates a genuine source to undertake a malicious operation at an Interex Conference in 1987. However, there does not appear to be a definite understanding of phishing techniques. Phishing has its own set of terminology that appears regularly in the literature. Mass, spam, and blanket phishing are examples of such words. They all have characteristics in common, such as many messages sent, misleading targeted individuals, impersonation of a sender, and data collection via social engineering [11–13].

Studies on phishing attacks' occurrence and success rate are conducted regularly. Overall, they all show an increasing tendency of phishing attacks [5, 7, 11, 14].

Previous studies have suggested that users are more prone to phishing if they are solicited by known entities in more targeted e-mails [14–17].

2.2. Phishing Susceptibility and Demography

Studies have been conducted to measure the relationship between demographic factors and phishing susceptibilities [14, 17–19] and to identify factors that predict phishing susceptibility [20].

Younger students presented themselves as more vulnerable in Jagatic et al.'s study, in which females became victims in 77% of cases, while males' proportion was 65% [14]. This study was performed on 487 selected students from Indiana University aged 18–24 years. On the other hand, this study was unique because it used personal information acquired from social media to send phishing messages to a target pretending to be a known friend.

Sheng et al. performed a role-play survey shared with 1001 respondents (only 29% of them were students) to learn more about the relationship between demographics and phishing susceptibility. Their results showed that females were more prone to phishing than males [18]. This is because females had less technical training and technical knowledge than males. Another finding was that participants aged 18–25 years were more susceptible than other groups. The age category of 18–25 years corresponds to the approximate age of university students.

Researchers from Carnegie Mellon University explored different age groups in their empirical phishing experiment. The study was based on the sending of phishing e-mails to a group of 515 participants. Results showed that 62.3% of the users in the age group of 18–25 years fell prey to the phishing e-mail, while 41.1% of the users in the age group of 26 years or older were tricked similarly [21].

Hong et al. explored the behavioural, cognitive, and perceptual attributes that make individuals more vulnerable to phishing. Of 53 respondents, over 92% were somewhat defenceless towards phishing [22]. In this experiment, it was revealed that females were less likely to uncover phishing e-mails.

Diaz et al.'s study conducted in 2019, where phishing e-mails were sent to 450 uninformed students at the University of Maryland, resulted in 60% of participants clicking on the phished e-mail; however, the study discovered no significant correlation between gender and susceptibility [19].

In Broadhurst et al.'s quasi-experimental study, 138 students were exposed to fake e-mails to connect demographic factors to phishing susceptibility. However, no correlation was found. All the variables indicated that international and first-year students were deceived more significantly than domestic and later-year students [17].

Many studies have been conducted over the past few years, mostly based on role-play investigations. This setup allows researchers to assess the effectiveness of phishing attacks without undertaking real-world phishing tests. Users respond to questions using role-playing to examine a possible security situation. The preliminary findings are analysed and summarised to identify potential phishing victims [14, 18].

A controlled phishing experiment was also used, in which individuals were sent an actual phishing e-mail that directed them to a phishing website. The phishing website does not capture or keep any personal data. On the other hand, this website keeps track of the number of victims and perhaps their usernames. The information gathered can be used to measure user security awareness and, in the future, to improve security training [15, 19, 21, 23].

Although all previous studies focused on demography and susceptibility, they used different methods to find out results. None of the above studies explicitly focused on the gender of students and

susceptibility to phishing by using phishing simulation. The current study uses the latter technique and concentrates on finding whether females or males in a university setting are more prone to opening a phishing e-mail, and which type of e-mail is opened more often by sending a decoy e-mail to registered participants to help understand the current vulnerability of students of social studies.

3. Methods

In this section, details of the methodology of this research are presented.

3.1. Structural Overview

The aim is to create conditions similar to a real-life environment while maintaining secure surroundings for collaboration, privacy, and dignity in research. We opted for a phishing simulation campaign in which realistic decoy e-mails were sent to students. At the same time, Google Analytics and SalesHandy helped us gather accurate data on the dangers of phishing on social studies faculty – two phishing e-mails were used – one generic, although adjusted to the current situation, and another targeted (spear). The first phase comprised obtaining a list of target identities to experiment on; and the second phase comprised preparing a technical background for an experiment. This was followed by sending decoy phishing e-mails and gathering data.

3.2. Data Collection

The first step was to assemble participants. Participants were not chosen randomly, as mentioned in the literature, but voluntarily through a registration form. The registration form consisted of questions on demographic information (age, gender, studies, year of studies, and language of their studies) and an informed consent. With this, participants had the chance to learn the purpose, benefits, and risks before deciding or declining to participate in the study. Crucial to the experiment was soliciting university e-mail, which was later used as an entrance to complete research. Students were approached in November 2021 through the social network Facebook and Discord; they were able to register until the end of the year. Responses were collected through E analyzer, a data-gathering and processing platform. Initially, 101 students registered. The number reduced to 68 due to incomplete data in some cases. After collection of all data, participants were assigned numbers to anonymise their identity and keep an overview

of the results. The Faculty of Social Studies at Masaryk University had 2804 students. With standard statistical technique [24], it was determined that a sample size of 68 students was applicable for a confidence level of 95% and with $\pm 11.33\%$ margin of error. For characteristics of the whole tested group, see Table 1.

3.3. Phishing Web Creation

The following step was to prepare a technological background to capture all feedback. To execute practical experiments, a functional website was needed, ideally similar to a faculty website that somehow counts the activity. For that purpose, the decision was to use the framework Django, a tool developed in the programming language Python. Because the only functionality requested of

Table 1. Characteristics of Sample.

Characteristic	Participants (N = 68)
Gender	
Male	44%
Female	54%
Others	1.5%
Age (years)	
<21	31
21–25	60
26–31	6
>31	3
Studies	
Environmental Studies	1.5
International Relations	40
Media Studies	7
Political Science	9
Psychology	7
Social Policy and Social Work	7
Sociology	7
Language	
Czech	90
English	10

the website was to appear legitimately and measure visits, a default Django project was created.

The consequent step was obtaining the visual side – which was achieved by using a web browser tool to view the HTML source code of the MUNI newsletter. This source code was copied, slightly adjusted, and set as a visual for the homepage of the Django project. We created one more page to count the hits of a targeted e-mail. The appearance of this page was not significant, so it contained just some simple HTML structures with the statement announcing that the visitor has been phished, it was a part of the experiment, and two useful links to relevant sources: one for the NÚKIB¹ website, and another for MUNI security.

For the last requirement, counting visitors, Django extension called Django-hit count was used. This extension counts webpage hits by analysing the requests sent – website traffic. Later on, it was found that it did not work as needed, so this option was abandoned, and it was decided to look for other options.

The Google Analytics tool was used for the purpose of this experiment. A Google token was generated to make it operative, and HTML to the page's source code was added. Besides counting the visitors, Google Analytics provided us with much additional helpful information about them, such as operating system, whether participants used mobile or desktop access, and the browser.

The Django project was then ready to be deployed online. Heroku hosting was used for that purpose because it provided simple free hosting for projects written in Django. Heroku also allowed the use of custom domain names, which were essential for the success of this experiment. Because it allowed adjusting the domain, it looked more similar to the MUNI domain (*muni.cz*).

The domain name we chose for the experiment was muninewsletter.cz, for two simple reasons: it looks identical to MUNI, and it enabled the use of social engineering. The information was obtained to make it look like a credible institution. After doing market research online to find the best offer for this domain, we opted to go with godaddy.com. It simultaneously created an e-mail for this domain suited for the usage. After purchasing the domain name, the only thing remaining was to set it to redirect to Heroku. That was achieved by setting nameservers at godaddy.com to redirect to Heroku nameservers, which then directed the user to our project.

1——NÚKIB is National Cyber and Information Security Agency in Czech Republic.

3.4. E-mail Setup

Creating a functionalised e-mail to store all responses was the next step when the website was set up and started working. For this purpose, a tool called SalesHandy was used. The e-mail address was chosen to be as similar as possible to the original. The e-mail address from which the university sends newsletter e-mails is studenti@newsletter.muni.cz. The e-mail address used for this experiment was studenti@newslettermuni.cz. The difference was in one dot. This method is called link manipulation; it is a technical disguise. The link is slightly altered to make the user believe it more and then redirects to the phisher's website.

After SalesHandy was connected to the e-mail address studenti@newslettermuni.cz, it enabled sending e-mails with tracking and planning the e-mail campaign. The most significant features of this tool were showing who opened, replied, and clicked on the link in the sent e-mail. This facilitated recording participants' behaviour after the decoy e-mail was sent.

3.5. Phishing E-mail Design

Phishing e-mails were inspired by the phishing archive of Berkeley University of California [25] to copy the usual visualities that real phishing e-mails in the university environment have. Social engineers use different techniques intending to be successful. Phishers are getting more sophisticated; phishing attacks incorporate greater details and context to become more effective and, therefore, more perilous for society [14]. Thus, both e-mails were written in the Czech language, because most registered respondents studied the Czech programme, and the main goal was to make it look real.

Because of that, with the first e-mail, we tried to be as precise as possible. For the first e-mail, the generic one, the decision was to copy the student's newsletter.

We used e-mail spoofing, where information from a section of the e-mail was falsified, making it appear as if it was coming from a legitimate source – Masaryk University. The second approach is website cloning; with this technique, we copied a legitimate website and an e-mail of the student newsletter and tried to deceive students into clicking on the link. These fake sites usually trick individuals into entering personally identifiable information (PII) or login credentials or attacking directly. For a higher click rate, the current situation was used. Specifically, students were presented

with an e-mail in which they could find more information on how Masaryk University is helping with the conflict in Ukraine. This topic was chosen because it is presently happening and attacks human emotions, which is one of the preconditions for successful phishing [9]. After clicking on the link, participants were redirected to the website, which looked like the webpage of Masaryk University but had spelling mistakes. This created space for conscious individuals to report this situation to MUNI IT team. The purpose of the second e-mail was to be more personal; hence, copying of an e-mail which announces the receiving of a document in the information system of Masaryk University. This e-mail was sent with spelling errors, and the link, <http://www.newslettermuni.cz/outside/>, did not match the e-mail's subject.

3.6. The Realisation of Experiments

Tryouts were executed before completing the first experiments to ensure that sent e-mails would not be delivered to spam.

The first e-mail was sent out on 2 March 2022. Two days later, the second e-mail type was sent on 4 March 2022. Two-month delay after collecting primary data was due to the waiting period, which was supposed to gain time to prepare the experiment's technical background and ensure that participants would not have a fresh memory of signing up for the experiment. E-mails were sent during the campaign, which ensured the delivery of e-mails at approximately the same time. Two days after the last e-mail was sent, the data was downloaded and converted to the .xlsx format for further analysis.

4. Results

Susceptibility is not homogenous among internet users; many factors influence individuals' decision-making and online behaviour. The present study seeks to determine the profile of a student most vulnerable to phishing and, based on results from previous research, confirm whether male or female students of FSS MUNI SCI are more susceptible to be victimised by phishing e-mails [14, 17–19]. The following text presents general observations of this study, followed by a comparison of the results from phishing susceptibility to two types of e-mails. In this research, falling for phishing is defined as clicking on the link in the e-mail, according to the research which was published in 2021 [26]. The distinction is made between not opening an e-mail, opening an e-mail and clicking on the link in the e-mail. The phishing campaign and collection of the responses lasted the first forty-eight hours after delivering the e-mail.

4.1. General Observations

Altogether, 136 e-mails were sent. Participants were most susceptible to the spear-phishing e-mail, which announced the delivery of a document to the IS of MUNI. This e-mail was tailored for the students of Masaryk University because it informs the recipient from the second-person point of view. Of all participants, 74% opened this e-mail, and 96% of those who opened it also clicked on the link. This was noticeably different in the case of the first e-mail, which was opened by 34% of respondents, and on the link clicked by 52% of those who had opened the e-mail. Overall, there appeared to be an increasing trend concerning scam and scam susceptibility in normalised proportions, with increasing success for more individualised and tailored scam rather than the generic one.

4.2. The First E-mail

The first e-mail was not opened by 66% of respondents. 34% of participants opened the e-mail, and 52% also clicked on the link contained in the e-mail. This e-mail aimed to be general but slightly adjusted for the attention of university students, so the e-mail domain fits the perspective. Regarding the male-female ratio, of 45 people who did not open the e-mail, 18 were males, 26 were females and one other. While this first e-mail was mainly ignored by females, the ratio was equivalent when opening the e-mail. Of 23 people who opened the e-mail, 12 were males and 11 were females. More females clicked on the link, but the difference was minimal; the male-female ratio was 5:7. This thesis focuses on 'male's and 'female's susceptibility to phishing; however, participants marked other demographic information in the registration form. For the whole list of characteristics, see Table 2. Data that were insignificant due to the low number of responses captured are excluded from the table. The success rate of this first e-mail was 18%.

4.3. The Second E-mail

The second e-mail brought different results. Only 18 participants did not open the e-mail, while 50 users opened it. From that, 48 people clicked on the link in the e-mail, making for a 96% clicking rate. In the case of the second e-mail, two persons alerted the CSIRT² MUNI team. Of 18 people who did not open the e-mail, 7 were males, and 11 were females. The e-mail was opened by 19 male respondents, 30 female respondents and one other. Further, 17 male and 30 female respondents clicked on the link, showing higher susceptibility to phishing in females. The sample of

2——CSIRT stands for Computer Security Incident Response Team, and it handles security incidents on computer networks. This type of group is usually associated with a specific region or organisation; in this case, the Masaryk University.

Table 2. Characteristics of the First E-mail.

Characteristics	Didn't open the e-mail (N = 45)	Opened the e-mail (N = 23)	Clicked on the link (N = 12)
Males	18	12	5
Females	26	11	7
Others	1	0	0
International Relations	18	9	4
Psychology	11	8	5
Political Science	6	1	0
>21 years	14	7	2
21–25 years	26	15	10
2nd	12	3	1
3rd	12	7	5
4th	11	5	4

respondents consists of a more significant proportion of females than males, and the reason for this is that it reflects more female students at the FSS; the male–female ratio was 30:37. The success rate of the second e-mail was 71%.

After clicking on the link, participants were directed to the page announcing that they were phished and linked to useful links to learn more about phishing attacks from the NÚKIB or MUNI security team.

The last e-mail of this type was sent on 4 March 2022 at 22:15. Approximately 24 h later, on 5 March 2022, respondents numbered 16 and 54 started a debate on the suspicious e-mail on the FSS virtual campus on discord. Participants discussed whether it was part of a training or a real security threat. After exchange of short messages, they concluded that the best would be to report it to the CSIRT MUNI. And so they did; both participants communicated this information to the relevant team, who told them that this was part of a research for thesis. For further data concerning the second e-mail, see Table 3.

4.4. Comparison

From the results listed above, it is clear that the second targeted e-mail was more successful; however, what was the difference? We examined the effect of gender on participants to

Table 3. Characteristics of the Second E-mail.

Characteristics	Didn't open the e-mail (N = 18)	Opened the e-mail (N = 50)	Clicked on the link (N = 48)
Males	7	19	17
Females	11	30	30
Others	0	1	1
International Relations	8	19	18
Psychology	4	15	14
Political Science	2	5	5
<21 years	6	16	14
21–25 years	9	32	32
1st	4	10	9
2nd	4	11	10
3rd	4	15	15
4th	6	10	10

see whether gender differences exist in responding to phishing susceptibility.

In the case of the first e-mail, the opening of an e-mail was comparable between genders. The results showed that 12 males and 11 females opened it.

Numbers almost doubled when it came to the targeted e-mail. The second e-mail was opened more times by females, even though the number of participants in both groups was roughly the same (30 males and 37 females were registered for this study).

It was found that the first type of phishing attack equally deceived female and male subjects. However, in the second type of phishing, almost 63% were female compared to 36% male victims, which was in accordance with the study of Jagatic et al. [14] and Sheng et al. [18], where the authors found that females were more susceptible to the spear-phishing risk.

A low percentage of subjects clicking on the link (18%) suggests that the more targeted the e-mail, the more significant the threat. From the second e-mail, it was clear that males were less susceptible to falling prey to phishing attacks than females. The results also indicated that females were more likely to click on phishing links.

4.5. Google Analytics – Profile of the User

Google Analytics provides additional information about the operating system and the browser through which users accessed the phishing site. This information helped to more accurately define the user profile of those who were victimised by phishing in this research and provided a framework for developing a new hypothesis in future research related to the factors that make users vulnerable to phishing attacks.

Because the first e-mail demonstrated a success rate of 18% for susceptibility to phishing, to create a profile of a student of FSS susceptible to phishing, the data obtained by the second decoy e-mail was used.

Females clicked on the link in the e-mail with a ratio of 30:17, thus making them user's first attribute. The highest click rate was in the age category of 21–25 years. However, this may be negligible due to the disproportion of the sample in this category. Across different university year groups, the sample was divided comparably. Students reached the highest susceptibility in the third year. From the perspective of studies, the highest number occurred for students of International Relations, followed by students of Psychology. Provided by Google Analytics, most users used Windows as their operating system and entered the web page from their desktop, specifically from the Google Chrome browser. Table 4 summarises all the factors connected with susceptibility to phishing attacks in the present research.

5. Discussion

According to the literature review, the specificity of scams may influence phishing attack susceptibility; that is, people are more likely to be deceived by scams tailored to their specific circumstances

Table 4. Profile of a Highly Susceptible User.

	Highly susceptible
Age (years)	21–25
Gender	Females
Education	International relation
Year of studies	3rd
Operating system	Windows
Desktop/mobile	Desktop
Browser	Google chrome

than scams with generic content. We used two scam types – generic and spear-phishing – to check whether respondents were more vulnerable to spear-phishing attempts than generic ones. This logic was confirmed. More than twice as many people opened the more targeted e-mail; 23 participants opened the generic e-mail, while the targeted phishing e-mail was opened by 50.

In the present study, the success rate was 18% for the first e-mail and 71% for the second e-mail, which is considerably closer to the type of phishing where e-mails were constructed based on gathered information.

The high success rate in the second e-mail indicates that students are more susceptible to targeted phishing than the generic one. This number is alarming but not unusual among university students. The success rate was comparable to the results of a study done by Jagatic et al. [14], which had a success rate of 72%. However, this high number opens a space for discussing basic cybersecurity knowledge among university students because they represent a highly vulnerable group.

Results suggest that the more susceptible gender to targeted phishing e-mails is females because the clicking rate in their case was 81%, while 63% males clicked on the link in the second phishing e-mail.

A real-life fraud experiment on human subjects was witnessed in this study, with highly valuable ethical implications. How can one learn about students' sensitivity to phishing without them knowing but keeping it in a natural setting? In this experiment, this was solved through a signed registration form. However, this ethical issue was at the expense of a more significant number of participants and also the moment of surprise, even though we waited for 2 months for the preparation of experiment post-registration.

This article, however, had several limitations. The first was the insufficiency of the sample size for generalisation. The final number of respondents was 68, as many had to be excluded because of insufficient information. Hence, a limitation in presenting the pattern of findings and analysis. Because of the small sample size, the scope of analysis was also limited. Even though it was not significant in the case of this study, this could be an opportunity for future researchers in this area because, as stated before, the number of clicking on the links was alarming. Statistics were partly collected manually, creating space for human error because of accidental occurrence of miscounting.

One of the aspects which helped this research was the usage of its own domain. This raised the overall level of legitimacy of e-mails. It also allowed to measure users' susceptibility to higher-level phishing attacks, requiring higher understanding and awareness to fall victims.

Future research could apply a more extensive phishing simulation to determine the variables influencing students' scam susceptibility. Understanding the factors that influence phishing susceptibility could help with customised cybersecurity education, thereby protecting against phishing and other forms of cybercrime.

6. Conclusions

In conclusion, this article presented the design and results of two phishing campaigns conducted among students of the Faculty of Social Studies at Masaryk University. Through a phishing campaign simulation using e-mails, the practical study enabled a deeper investigation of the phenomenon of phishing at universities, providing insight into the susceptibility of different genders. Based on the obtained click rate percentage, more cybersecurity education and awareness are required.

Results from phishing simulations indicate that students are prone to be victims of targeted phishing to a much greater extent than generic phishing e-mail, which does not compel action. Females opened and clicked on the phishing e-mails almost twice as often as males. According to the findings, phishing assaults are still one of the most severe threats to individuals and institutions. The phishing cycle is mainly driven by human interaction. Phishers frequently exploit human weakness, increasing the possibility of victimisation by phishing. Despite the limitations of this work, we consider it beneficial for a better understanding of the issues and future research. Exploring phishing threats and vulnerabilities in a university setting is especially crucial because everyone, employees and students alike, is accountable for handling the institution's data. As the sophistication of phishing attempts enhances, the likelihood of a university being targeted also increases. We can personalise focused prevention for such groups if we conduct a study and determine the most vulnerable groups.

Acknowledgements

At this point, the author would like to thank the people without whom the motivation and creation of this article would not have been possible: her parents, her siblings, Emma and Daniel,

and her long-time friends, Lucie and Samuel, who have always remained by her side.

References

- [1] M. Roser, H. Ritchie, E. Ortiz-Ospina. (2021). *Internet: Our world in data*. [Online]. Available: <https://ourworldindata.org/internet>. [Accessed: Mar. 28, 2022].
- [2] S. Venkatesha, K.R. Reddy, B.R. Chandavarkar, "Social engineering attacks during the COVID-19 pandemic," *SN Computer Science*, vol. 2, no. 2, p. 78, 2021, doi: [10.1007/s42979-020-00443-1](https://doi.org/10.1007/s42979-020-00443-1).
- [3] IBM. (2014). *IBM security services 2014 cyber security intelligence index: Analysis of cyber attack and incident data from IBM's worldwide security operations*. [Online]. Available: <https://i.crn.com/sites/default/files/ckfinderimages/userfiles/images/crn/custom/IBMSecurityServices2014.PDF>. [Accessed: Apr. 26, 2022].
- [4] F. Salahdine, N. Kaabouch, "Social engineering attacks: A survey," *Future Internet*, vol. 11, no. 4, p. 89, 2019, doi: [10.3390/fi11040089](https://doi.org/10.3390/fi11040089).
- [5] G. Bassett, C.D. Hylender, P. Langlois, A. Pinto, S. Widup. (2021). *Data breach investigations report*, DBIR. [Online]. Available: <https://www.verizon.com/business/resources/reports/2021/2021-data-breach-investigations-report.pdf>. [Accessed: Mar. 28, 2022].
- [6] P.W. Singer, A. Friedman, *Cybersecurity and cyberwar: What everyone needs to know*. New York, NY: Oxford University Press, 2014.
- [7] Anti-Phishing Working Group. (2022). *Phishing activity trends report: Unifying the global response to cybercrime 4th quarter 2021*. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q4_2021.pdf. [Accessed: Mar. 28, 2022].
- [8] M. Wagner. (2006). *Who's phishing for your students?* ZDNet. [Online]. Available: <https://www.zdnet.com/article/whos-phishing-for-your-students/>. [Accessed: Mar. 28, 2022].
- [9] Z. Alkhalil, C. Hewage, L. Nawaf, I. Khan, "Phishing attacks: A recent comprehensive study and a new anatomy," *Frontiers in Computer Science*, vol. 3, Mar. 2021, doi: [10.3389/fcomp.2021.563060](https://doi.org/10.3389/fcomp.2021.563060).
- [10] J.L. Bailey, B.K. Jensen, R.B. Mitchell, "Analysis of student vulnerabilities to phishing," Learning from the past & charting the future of the discipline. Proceedings of the fourteenth Americas conference on information systems, AMCIS 2008, Toronto, ON, Canada, Aug. 14–17, 2008. [Online]. Available: https://www.researchgate.net/publication/220891604_Analysis_of_Student_Vulnerabilities_to_Phishing. [Accessed: Apr. 03, 2022].
- [11] R. Heartfield, G. Loukas, D. Gan, "You are probably not the weakest link: Towards practical prediction of susceptibility to semantic social engineering attacks," *IEEE Access*, vol. 4, pp. 6910–6928, 2016, doi: [10.1109/ACCESS.2016.2616285](https://doi.org/10.1109/ACCESS.2016.2616285).
- [12] J. Hong, "The state of phishing attacks," *Communications of the ACM*, vol. 55, no. 1, pp. 74–81, 2012, doi: [10.1145/2063176.2063197](https://doi.org/10.1145/2063176.2063197).
- [13] E.E.H. Lastdrager, "Achieving a consensual definition of phishing based on a systematic review of the literature," *Crime Science*, vol. 3, no. 1, pp. 1–10, 2014, doi: [10.1186/s40163-014-0009-y](https://doi.org/10.1186/s40163-014-0009-y).

- [14] T.N. Jagatic, N.A. Johnson, M. Jakobsson, F. Menczer, "Social phishing," *Communications of the ACM*, vol. 50, no. 10, pp. 94–100, 2007, doi: [10.1145/1290958.1290968](https://doi.org/10.1145/1290958.1290968).
- [15] R.C. Dodge, C. Carver, A.J. Ferguson, "Phishing for user security awareness," vol. 26, no. 1, pp. 73–80, 2007, doi: [10.1016/j.cose.2006.10.009](https://doi.org/10.1016/j.cose.2006.10.009).
- [16] E.D. Frauenstein, "An investigation into students responses to various phishing emails and other phishing-related behaviours," 17th International Conference, ISSA 2018, Pretoria, South Africa, August 15–16, 2018, Revised Selected Papers (Communications in Computer and Information Science Book 973). New York, NY: Springer, 2019, pp. 44–59, doi: [10.1007/978-3-030-11407-7_4](https://doi.org/10.1007/978-3-030-11407-7_4).
- [17] R. Broadhurst, K. Skinner, N. Sifniotis, B. Matamoros-Macias, "Cybercrime risks in a university student community," *SSRN Electronic Journal*, vol. 2a, no. 1a, pp. 5–10, 2020, doi: [10.2139/ssrn.3176319](https://doi.org/10.2139/ssrn.3176319).
- [18] S. Sheng, M. Holbrook, P. Kumaraguru, L.F. Cranor, J. Downs, "Who falls for phish?," Proceedings of the 28th International Conference on Human Factors in Computing Systems, Atlanta, GA: ACM Press, USA, 2010, pp. 373–382, doi: [10.1145/1753326.1753383](https://doi.org/10.1145/1753326.1753383).
- [19] A. Diaz, A.T. Sherman, A. Joshi, "Phishing in an academic community: A study of user susceptibility and behavior," *Cryptologia*, vol. 44, no. 1, pp. 53–67, 2020, doi: [10.1080/01611194.2019.1623343](https://doi.org/10.1080/01611194.2019.1623343).
- [20] M.K.K. Tornblad, K.S. Jones, A.S. Namin, J. Choi, "Characteristics that predict phishing susceptibility: A review," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, pp. 938–942, 2021, doi: [10.1177/1071181321651330](https://doi.org/10.1177/1071181321651330).
- [21] P. Kumaraguru, S. Sheng, A. Acquisti, L.F. Cranor, J. Hong, "Teaching Johnny not to fall for phish," *ACM Transactions on Internet Technology*, vol. 10, no. 2, pp. 1–31, 2010, doi: [10.1145/1754393.1754396](https://doi.org/10.1145/1754393.1754396).
- [22] K.W. Hong, C.M. Kelley, R. Tembe, E. Murphy-Hill, C.B. Mayhorn, "Keeping up with the Joneses," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 57, no. 1, pp. 1012–1016, 2013, doi: [10.1177/1541931213571226](https://doi.org/10.1177/1541931213571226).
- [23] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M.A. Blair, T. Pham, "School of phish," Proceedings of the 5th symposium on usable privacy and security (SOUPS '09). New York, NY: ACM, 2009, doi: [10.1145/1572532.1572536](https://doi.org/10.1145/1572532.1572536).
- [24] R.V. Krejcie, D.W. Morgan, "Determining sample size for research activities," *Educational and Psychological Measurement*, vol. 30, no. 3, pp. 607–610, 1970, doi: [10.1177/001316447003000308](https://doi.org/10.1177/001316447003000308).
- [25] Berkeley Information Security Office. (2023). *Phishing examples archive* [Online]. Available: <https://security.berkeley.edu/education-awareness/phishing/phishing-examples-archive>. [Accessed: Apr. 04, 2023].
- [26] A. Jayatilaka, N.A. Gamagedara Arachchilage, M.A. Babar, "Falling for phishing: an empirical investigation into people's email response behaviors," Proceedings of the 42nd international conference on information systems (ICIS), 2021, Austin, TX. Atlanta, GA: Association for Information Systems, pp. 1–6, 2021, doi: [10.48550/arXiv.2108.04766](https://doi.org/10.48550/arXiv.2108.04766).

Deepfake Influence Tactics through the Lens of Cialdini's Principles: Case Studies and the DEEP FRAME Tool Proposal

Received: 17.09.2024

Accepted: 13.12.2024

Published: 30.12.2024

Cite this article as:
P. Zegarow, E. Bartuzi, "Deepfake influence tactics through the lens of Cialdini's principles: Case studies and the DEEP FRAME tool proposal," ACIG, vol. 3, no. 2, 2024, pp. 286–302. DOI: 10.60097/ACIG/201147

Corresponding author:
Pawel Zegarow, Strategy and Development of Cyberspace Security Team, NASK – National Research Institute, Poland; E-mail: pawel.zegarow@nask.pl
 0009-0006-5421-8656

Copyright:
Some rights reserved (CC-BY):
Pawel Zegarow,
Ewelina Bartuzi
Publisher NASK

Pawel Zegarow | Strategy and Development of Cyberspace Security Team, NASK – National Research Institute, Poland | ORCID: 0009-0006-5421-8656

Ewelina Bartuzi | Audiovisual Analysis and Biometric Systems Department, NASK – National Research Institute, Poland | ORCID: 0000-0001-6245-2908

Abstract

The advancement of artificial intelligence (AI) has introduced both significant opportunities and challenges, with deepfake technology exemplifying the dual nature of AI's impact. On the one hand, it enables innovative applications; on the other, it poses severe ethical and security risks. Deepfakes exploit human psychological vulnerabilities to manipulate perceptions, emotions, and behaviours, raising concerns about the public's ability to distinguish authentic content from manipulated material. This study examines the methods of influence embedded in deepfake content through the lens of Robert Cialdini's six principles of persuasion. By systematically analysing how these mechanisms are employed in deepfakes, the research highlights their persuasive impact on human behaviour, particularly in scenarios such as financial fraud. To address the challenges posed by deepfake technology, this study introduces DEEP FRAME, an original tool for systematically recording and analysing deepfake content. DEEP FRAME integrates technical and psychological analysis, enabling the identification of technological characteristics and manipulation strategies embedded within deepfakes. The findings underscore the need for a holistic and interdisciplinary approach that combines technological



innovation, psychological insights, and legal frameworks to counter the growing threat of deepfakes.

Keywords

social engineering, cybersecurity, cyberpsychology, deepfake, DEEP FRAME tool

1. Introduction

While artificial intelligence (AI) has the potential to revolutionise key aspects of human life – from work and education to health and security – its development also introduces significant ethical and societal risks. One of these is the emergence of deepfake technology, identified by researchers at University College London in 2020 as one of the most dangerous AI-enabled crimes due to its potential for malicious use [1].

Deepfakes actually pose a significant challenge to researchers and practitioners seeking effective strategies to protect individuals and societies from manipulation. The potential of deepfakes to influence perception and decision-making raises serious concerns about information integrity across a variety of sectors, including politics, media, and the private domain. Despite the critical nature of this problem, existing research has mainly focused on technical aspects, such as detecting fake content. Therefore, our work focuses on an interdisciplinary approach, combining technological and psychological knowledge about the mechanisms of persuasion used in deepfakes. Additionally, we propose the DEEP FRAME tool for the systematic analysis of deepfakes, which takes into account both technical and psychological aspects.

The term 'deepfake' is derived from English and combines two words: 'deep learning' and 'fake'. It refers to a technology that uses AI to generate videos, images, and sounds so realistic that distinguishing it from authentic material becomes challenging. The term 'deepfake' was first used at the end of 2017 by an anonymous Reddit user operating under the pseudonym 'deepfakes'. This individual utilised deep learning methods to create video content in which the faces of performers in adult films were replaced with those of recognisable public figures [1–3].

1.1. The Unexpected Challenges of Detecting Deepfakes

Until recently, content generated by AI was characterised by relatively low quality and contained easily identifiable errors,

such as distorted facial features, unnaturally rendered skin, or hands depicted with more than five fingers. However, advancements in machine learning techniques and the increasing computational power of modern systems have made contemporary deepfakes nearly indistinguishable from authentic content, exhibiting an exceptionally high level of realism.

It is important to note that not all deepfakes produced today achieve hyperrealistic quality, but many of these videos are consumed on mobile devices, where smaller screen sizes and lower resolution settings can mask imperfections and make distinguishing them from real content significantly more challenging.

Detection of deepfakes is a complex issue. Research shows that human ability to recognise deepfakes varies considerably, with detection accuracy ranging from 57% to 89%. This suggests that even in the most optimistic scenario, individuals fail to identify 11 out of every 100 deepfakes, while in the most pessimistic case, as many as 43 out of 100 deepfakes go undetected [4].

1.2. Deepfake Technology as a Tool for Cybercriminal Activity

Although the online dissemination of false content is not a novel phenomenon, the advent of deepfake technology has enabled cybercriminals to engage in malicious activities on an unprecedented scale. Deepfakes are employed in a range of criminal endeavours, including extortion, reputational damage, the manipulation of political processes, disinformation, and financial frauds. According to the assessment conducted by the authors of the publication 'AI-enabled future crime', deepfake technology may inflict the most substantial harm and yield the greatest potential profits for cybercriminals [5].

Cybercriminals frequently integrate deepfake technology with social engineering strategies to enhance the effectiveness of their activities. They exploit various human psychological vulnerabilities, such as the ease of eliciting trust, the propensity to act under time pressure, and susceptibility to suggestion. By doing so, cybercriminals can construct deceptive narratives and influence the behaviours of potential victims, ultimately aiming to gain access to sensitive information.

Deepfake technology poses threats not only to individual users but also to businesses, government institutions, and international organisations. By exploiting human beings' natural tendency to

trust the authenticity of perceived images and sounds, cybercriminals destabilise cognitive processes, affecting emotions, attitudes, and perceptions of reality. As a result, victims of such manipulation often make irrational decisions and engage in behaviours based on false premises, potentially leading to severe personal and professional repercussions.

In March 2022, as reported by the *Reuters* news agency, a deepfake of Ukrainian President Volodymyr Zelensky appeared on the social media platform X, calling on Ukrainian soldiers to lay down their weapons. It can be inferred that the purpose of this deepfake was to undermine morale and create confusion during an active armed conflict. Although the footage was almost immediately identified as inauthentic, its deceptive nature was not immediately apparent to all viewers – particularly older people or those less familiar with technological advances. This example illustrates how dangerous the spread of false content on social media can become in times of armed conflict and highlights the serious social and political repercussions that can result.

In recent times, the Russian Federation's production and large-scale dissemination of deepfake content has become increasingly intense. Merely one day following the terrorist attack of 22 March 2024, a fabricated video emerged in which the image of Oleksiy Danilov, Secretary of the National Security and Defense Council of Ukraine, was integrated into a format resembling a professional television interview. This footage served as part of a broader disinformation campaign aimed at attributing responsibility for the incident to Ukraine. To create this material, archived footage from 16 March 2024 – originally depicting Kyrylo Budanov, the head of the Main Directorate of Intelligence of the Ministry of Defense – was repurposed by substituting his likeness. Despite propagandists' efforts, the quality of the fabricated material remained low. Advanced voice cloning tools and lip sync techniques were employed to synchronise mouth movements with the manipulated statements; however, the final result was far from seamless. Visual distortions, unnatural synchronisation of lip movements with speech, and notable blurring – especially around the neck and mouth – remained discernible to vigilant observers. It must be acknowledged, however, that not all audiences are capable of detecting such subtle signs of interference. A previous attempt at using a similar disinformation strategy was recorded in October 2022, involving voice cloning to gain access to confidential information from the drone manufacturer Bayraktar. In this instance, perpetrators impersonated Ukraine's Prime Minister, Denys Shmyhal,

but their attempt was thwarted. This underscores the need for continuous improvement in methods of detection and countering such threats.

Recently, there has been a marked intensification in the use of fraudulent investment advertisements as tools for extracting personal data and financial resources. Such content appears on social media platforms, websites, and even as sponsored advertising materials. A defining characteristic of these schemes is the promise of rapid profits with minimal risk, often reinforced by the use of recognisable figures from the worlds of politics, business, entertainment, or finance.

Fraudsters increasingly produce complex audiovisual materials whose aesthetics and format resemble those of news programmes. They incorporate the likenesses of well-known presenters and journalists, who appear to endorse ‘investment opportunities’ or prompt viewers to take specific actions, such as clicking on a link or downloading an application.

According to information published in the British newspaper *The Guardian*, in early February 2024, the Hong Kong police reported that cybercriminals had employed deepfake technology to steal nearly £20 million. An investigation was launched following a report from an employee of a British company operating a branch in China, who informed the police that she had been coerced into transferring a significant sum of money into bank accounts designated by individuals posing as high-ranking company officials. Prior to executing the transfer, the employee had participated in a videoconference with the chief financial officer and other members of the management team. The investigation subsequently revealed that the individuals participating in this meeting were generated by AI [6].

Cybercriminals produce a diverse array of materials, each tailored to distinct target audiences. To promote fictitious investment ventures, they frequently use the likenesses of politicians and business leaders, intending to foster trust among individuals interested in traditional forms of investment. In contrast, content featuring celebrities and influencers are commonly deployed in advertising mobile applications – such as those simulating online casinos or games – primarily attracting younger audiences seeking entertainment. For older or unwell individuals, cybercriminals create deceptive ‘miracle’ drug or medical procedure advertisements, capitalising on the credibility associated with renowned physicians,

athletes, or religious figures. Among the widely publicised cases reported by the media are as follows:

- A 'gas pipeline investment' scam in which an elderly resident of Lower Silesia lost half a million PLN after believing promises of high returns on a purported investment project [7].
- A crime involving the misappropriation of funds under the guise of investing in Baltic Pipe gas pipeline shares. Victims, enticed by the reputation of this strategic energy project, ultimately lost their life savings [8].
- The case of a 35-year-old woman who invested approximately 150,000 PLN in a fraudulent crypto currency scheme. Exploiting her lack of experience and promising quick and guaranteed profits, the scammers induced significant financial losses [9].

All the above examples underscore the growing significance of fraudulent investment advertisements as tools employed by cybercriminals. Given the substantial social harm caused by such offenses, it is imperative to conduct in-depth research into the mechanisms governing the creation, distribution, and reception of this type of content, as well as to implement effective educational programmes and advanced technological measures. Such actions are crucial for reducing the scale of losses suffered by potential victims and for enhancing security within the digital environment.

It should be emphasised that deepfakes and social engineering are mutually reinforcing phenomena, giving rise to a new generation of threats characterised by a high level of technological sophistication and effectiveness in manipulating human behaviour. Unfortunately, conventional protection methods, which predominantly rely on technological safeguards, have proven insufficient. Consequently, countering deepfakes necessitates an interdisciplinary approach that integrates psychological, technological, and legal expertise.

2. Purpose

This study aimed to analyse the use of Cialdini's persuasion strategies in deepfake videos and present a DEEP FRAME – an original tool for recording and analysing deepfake content.

3. Methods

3.1. Deepfake Selection and Transcription

Given the objectives outlined in this study, a deepfake video observed on social media platforms in Poland between May

and July 2024 was selected to serve as the central case study. The sample was limited to Polish social media platforms to tailor the analysis to the local cultural and social context. The sample selection in our study was intentional. This video was purposefully chosen to ensure that it provided a diverse and comprehensive representation of deepfake-related content, encompassing a wide range of narrative structures, emotional triggers, and psychological attributes. Transcript analysis, therefore, incorporated both qualitative and quantitative methodologies, offering insights into the persuasive techniques at play, including mechanisms rooted in emotional appeal, authority, and social proof.

In the second stage, the researchers conducted the transcription of deepfake videos. The study employed automatic speech recognition (ASR) models to transcribe audio from deepfake videos. These models, often underpinned by large language models (LLMs), are specifically designed to accurately transcribe audio recordings into written text.

It should be emphasised that one of the primary challenges faced by ASR models is their limited ability to process noisy audio recordings. In cases of low-quality audio, where background noise, music, or other interferences are present, transcriptions may be incomplete or inaccurate. Under such conditions, models often generate extraneous elements in the text, such as phrases like 'subtitles sounds...' or other artifacts caused by misinterpretation of background noise. Another significant challenge arises when interpreting words that are mispronounced or articulated in an unusual manner. ASR systems tend to substitute such words with alternatives that better align with the surrounding context. This can result in distortions, particularly when proper nouns, technical terms, or slang expressions are replaced with incorrect equivalents.

Furthermore, transcriptions may contain repetitive words, sentences, or even entire speech segments. These repetitions often stem from the model's uncertainty regarding the interpretation of specific audio fragments or errors in the speech recognition algorithm. ASR models also struggle with recordings that feature shifts in accent or pronunciation. This issue becomes particularly pronounced when the speaker's intonation changes or when an accent characteristic of another language is introduced, such as Polish speech interspersed with Russian or English accents. In such scenarios, models may generate what are referred to as 'linguistic hallucinations,' introducing foreign language fragments into the transcription [10].

3.2. Cialdini's Persuasion Strategies

This study applied Robert Cialdini's six principles of social influence – reciprocity, commitment and consistency, social proof, authority, liking, and scarcity – to analyse deepfake [11]. The deepfake's transcript was analysed by a psychologist, who, drawing on their knowledge of influence techniques and professional experience, classified specific sections of the text as particularly persuasive. This assessment considered key persuasion mechanisms, such as authority, social proof, emotional appeal, and the principles of scarcity and reciprocity.

The principle of reciprocity refers to the innate human tendency to reciprocate benefits received from others, even in the absence of necessity. The analysis focused on identifying elements within deepfake videos that could evoke a sense of obligation or an inclination to reciprocate in the viewer. For instance, the videos may imply exclusivity in the presented information, potentially prompting viewers to reciprocate by further sharing the content.

The principle of commitment and consistency emphasises individuals' tendency to maintain alignment between their actions and decisions over time. The study examined whether the deepfakes employed techniques designed to prompt initial low-commitment actions, such as liking or sharing content, which could subsequently foster greater engagement.

The principle of social proof is based on the influence of others' behaviours on an individual's decision-making process, particularly in new or ambiguous situations, where people tend to follow the actions of others as a guide. The analysis assessed whether the deepfake materials incorporated elements, such as positive comments, or references to perceived broad social support, which could amplify the message by creating the impression that the stance presented is widely accepted and endorsed.

The authority principle is based on the tendency of individuals to place trust in and act upon information provided by perceived experts or leaders. The analysis examined whether the deepfake materials featured prominent figures, such as politicians, scientists, or opinion leaders, leveraging their perceived authority to enhance the credibility and impact of the message.

The liking principle refers to the idea that individuals are more likely to be persuaded by those they perceive as likable or who share similarities with them. The analysis investigated whether the deepfake

materials employed references to shared cultural and social values to strengthen their persuasive impact on recipients.

The principle of scarcity emphasises the perceived value of information by suggesting its limited availability. The study analysed whether the deepfake materials strategically employed techniques implying rarity, exclusivity, or urgency in the message, which could facilitate expedited decision-making or deepen viewer engagement through psychological pressure.

4. Results

4.1. Deepfake Analysis

4.1.1. Deepfake Transcript

You all know me. My name is Rafał Brzoska. I am a professional businessman and investor. Today is your lucky day. This page is available to only 100 people, and you are one of the few who will have the opportunity to make money and change your life. Only the most determined individuals will be able to achieve this. Of the 100 invited, only fifty of the most ambitious will take advantage of my offer. So let's get straight to the point. When I say this will change your life, I don't mean 2000 or 10,000 złotys. I mean an amount that will allow you to quit your job and go home. It's like early retirement or an additional pension – several times larger than your regular savings.

Before you leave this page thinking I'm a complete fool, wait a moment and listen to me. This isn't another video about someone trying to scam you out of your money, because I respect you and want to earn your trust. I won't make empty promises like everyone else. What's the difference between those scammers and me? First, I am Rafał Brzoska, and I don't need anything from you. I will provide you with proof that my project actually works. Promising you millions tomorrow is a lie, just like other empty promises, but four thousand PLN a day is absolutely achievable. Just do the math. Four thousand PLN a day equals 28,000 PLN a week or 1,96,000 PLN a month.

Now listen carefully. This video can only be viewed once. If you leave this page, you won't have another chance to return because your link will expire, as will your opportunity to make money. This has nothing to do with Forex stocks, financial pyramids, or any other nonsense you may see everywhere. I spend most of my life creating projects

that can improve the lives of every individual. The software we develop is better than all competitors thanks to its unique technologies and can be used on any computer or phone. Its unique AI-based analytical features allow it to stay ahead of market trends, ensuring exceptional success across all financial markets. The algorithm does everything for you. All you have to do is watch the results.

We tested our product on a small group of volunteers, each of whom earned over 24,000 PLN within the first week. I don't want to disappoint you, but you can only make 16,000 PLN, which is still a lot of money, isn't it? No, you don't need any special skills. If you're watching this video, it means your device supports this platform. You're probably wondering why I chose you and didn't keep such a unique algorithm for myself if it can generate such significant profits. The answer is simple. I decided to offer this program to 100 random users, but it turned out that only fifty of them would actually try to change their lives. I hope you are among those fifty people who want to become financially independent because you have the chance to try our algorithm for free and change your life.

In return, I ask you to write a short review so that people for whom our program will cost 4000 PLN a month know it really works. The more free clients I get now, the more paying customers I'll have once I start selling the program. But let's get back to the point. Imagine receiving 4000 PLN every day. You'll no longer have to worry about not having enough money to pay your rent. Taking a vacation several times a year? No problem. Buying a house and paying off all your debts? That's no issue. Providing your children with access to the best schools and travelling to the most luxurious destinations will certainly not be a problem. Imagine a life where you don't have to worry about anything.

The most important thing is time. If you start at the right moment, you can earn a lot of money. If you delay, you may end up like 99% of other people. Now the most important information. Pay close attention. To start using the program, you need to visit the website, enter the necessary details, and then your personal manager will contact you to answer all your questions and grant you access to the platform. From that moment, your life will change. You can call it a new life, and I'm sure you won't regret it.

The transcription demonstrates that persuasion mechanisms are systematically and deliberately employed, creating a cohesive and highly effective framework for persuasion. By integrating a nuanced combination of emotional and rational appeals, the deepfake strategically manipulates the audience, fostering a sense of trust, urgency, and a perceived necessity to act. This structured approach highlights the potential of deepfakes to exert significant influence on individual decision-making and behaviour, particularly within the context of digital environments. Table 1 provides a detailed analysis of the transcripts.

4.2. DEEP FRAME tool

In Appendix 1, we propose a DEEP FRAME tool designed for analysts that enables a comprehensive examination of deepfake content through an interdisciplinary approach. DEEP FRAME is a self-reported tool that includes a set of questions about both technical and psychological elements of deepfakes.

The proposed tool offers a wide range of benefits, serving as a valuable resource for interdisciplinary research on deepfake content by integrating both technological and psychological perspectives. The data and analyses it generates have the potential to advance significantly the development of more sophisticated algorithms for detecting manipulated content. Beyond its contributions to research and detection, the tool holds considerable promise for public education, as the insights it provides can support the design and implementation of effective awareness campaigns. Moreover, it strengthens our capacity to understand emerging threats and devise targeted strategies to mitigate their impact. By combining elements from computer science, psychology, and communication studies, it ensures a more thorough evaluation of the subtle and overt manipulations embedded in digital media.

The DEEP FRAME tool facilitates the systematic collection of knowledge about deepfakes. The collected data is categorised based on key parameters, such as the type of manipulation, the technologies used, and the context of publication. A technical module provides analysis of the quality of video and audio, potential artifacts of manipulation. A unique feature of the tool is its ability to conduct psychological evaluations of deepfake content. This module identifies persuasive techniques. Such analyses deepen our understanding of the manipulative mechanisms employed in deepfake content. The DEEP FRAME tool enables the monitoring of global trends in the use of deepfake technology, helping to identify emerging threats.

Table 1. Psychological analysis.

The principle of social influence	Examples	Frequency	Comments
The principle of reciprocity	<p>'I respect you and want to earn your trust'.</p> <p>'The more free clients I get now, the more paying customers I'll have once I start selling the program'.</p> <p>'...you have the chance to try our algorithm for free and change your life, in return, I ask you to write a short review...'</p>	3	Building commitment by offering a free program and expressing appreciation. Offering something for free may create a sense of obligation in the recipient to write a review in return or take action.
The principle of commitment and consistency	<p>'Before you leave this page thinking I'm a complete fool, wait a moment and listen to me'.</p> <p>'Only the most determined individuals will be able to achieve this'.</p> <p>'Of the 100 invited, only 50 of the most ambitious will take advantage of my offer'.</p> <p>'I hope you are among those 50 people who want to become financially independent'.</p>	4	Creating a sense of uniqueness and the need to act among the chosen ones. The recipient is gradually drawn into the process through initial commitments, such as qualifying for the 'most ambitious' and 'most determined' group. This principle emphasises the need for consistency in action – once the recipient has been selected, they should prove their determination by taking advantage of the offer.
The principle of social proof	<p>We tested our product on a small group of volunteers, each of whom earned over 24,000 PLN within the first week.</p> <p>'... and you are one of the few who will have the opportunity to make money and change your life'.</p>	2	Pointing to the success of other users and the elitism of the group. Social proof elements are visible through references to group tests and the success of other users. Emphasising the 'exceptionality' of the group of 100 people creates a sense of elitism, but also of universal support for this initiative.
The principle of authority	<p>'You all know me. My name is Rafał Brzoska. I am a professional businessman and investor'.</p> <p>'I am Rafał Brzoska, and I don't need anything from you'.</p> <p>'The algorithm does everything for you. All you have to do is watch the results'.</p>	3	Using Rafał Brzoska's image and technology as a source of credibility. The principle of authority is applied by referring to a person (businessman Rafał Brzoska) presented as an expert and emphasising the technological advantage of the algorithm.
The principle of liking	No clear manifestations of sympathy building	0	
The principle of scarcity	<p>'This page is available to only 100 people'.</p> <p>'This video can only be viewed once'.</p> <p>'If you leave this page, you won't have another chance to return because your link will expire, as will your opportunity to make money'.</p>	3	Time pressure and limited number of places as a motivator for action. The inaccessibility is strongly emphasised by the limited number of places, unique access to the site, and the need to make a decision immediately. This principle increases the pressure on the recipient to act quickly.

5. Conclusions

Artificial intelligence has the potential to revolutionise key aspects of human life; however, it also introduces profound ethical and societal challenges. Deepfake technology exemplifies this dual nature, offering innovative opportunities on one hand and unprecedented ethical and societal risks on the other. As highlighted in this study, deepfakes have become a sophisticated tool capable of undermining trust, spreading disinformation, and facilitating cyber-crime on a global scale.

The authors of this study argue that it is crucial to reserve the term 'deepfake' exclusively for materials explicitly designed to mislead, propagate disinformation, manipulate, cause harm, or discredit individuals. This is particularly relevant to malicious applications, such as financial frauds, disinformation campaigns, blackmail, or criminal impersonation. We are aware that this definition of the term 'deepfake' is narrow, but we intentionally adopt it to emphasise the unethical nature of deepfakes. While broader definitions may encompass a wide range of applications, including creative and harmless uses, our approach facilitates a clear distinction between the ethical and innovative uses of AI – those that foster creativity and positively impact the society – and practices that violate individual rights and erode public trust.

Deepfakes, when combined with social engineering techniques, exploit human psychological vulnerabilities, such as trust and urgency, to manipulate perceptions, emotions, and behaviours. Leveraging established principles of persuasion, such as those identified by Cialdini, they amplify the perceived credibility of messages and raise significant concerns about the ability of individuals and organisations to distinguish truth from deception in critical scenarios, including political campaigns, armed conflicts, and financial frauds.

The growing sophistication of deepfake technology, often rendering content indistinguishable from authentic material, poses severe challenges for detection. Even in cases where imperfections are present, the prevalent use of mobile devices for content consumption obscures subtle indicators of manipulation, further complicating detection efforts.

Countering deepfakes effectively necessitates an interdisciplinary approach that integrates psychological insights, advanced technological tools, and robust legal frameworks. Psychological research plays a pivotal role in elucidating how deepfakes influence human

behaviour while technological advancements improve detection capabilities. Simultaneously, legal measures must address regulatory gaps to ensure accountability for the misuse of such technologies.

This study introduces DEEP FRAME, an innovative tool designed to systematically record and analyse deepfake content. By integrating technical and psychological analysis, DEEP FRAME enables the collection of critical data, including technological characteristics, emotional impact, and manipulation patterns, fostering the development of more effective countermeasures. Additionally, the tool supports interdisciplinary collaboration by creating a comprehensive database to inform educational initiatives, policy-making, and technological advancements. The DEEP FRAME tool facilitates the systematic collection of knowledge about deepfakes. The collected data is categorised based on key parameters, such as the type of manipulation, the technologies used, and the context of publication. These categorisations help to reveal patterns and correlations, providing a structured basis for further analysis and tailored countermeasures. This tool enables the monitoring of global trends in the use of deepfake technology, helping to identify emerging threats.

Deepfakes represent a new generation of threats, combining technological sophistication with manipulative effectiveness. The findings of this study underscore the urgent need for a collaborative response that integrates technological innovation, psychological research, and legal regulation. Tools such as DEEP FRAME play a critical role in advancing these efforts, offering a comprehensive platform to analyse and mitigate the risks posed by this rapidly evolving technology. By addressing these challenges holistically, society can navigate the ethical and security dilemmas associated with AI and safeguard trust in the digital age.

5.1. Limitations

The study was limited to a single case, which may not reflect the full diversity of situations occurring in the deepfake phenomenon. The results may be difficult to generalise to a larger population or other contexts. The results depend on the quality of the data available on the case being studied. Short deepfakes that last, for example, for 1 minute will have significantly fewer persuasive fragments compared to longer materials that last, for example, for 3 minutes. However, the authors of the study plan to conduct

further in-depth research in the future on the phenomenon of using influence methods in deepfakes.

References

- [1] M. Caldwell, J. T. Andrews, T. Tanay, L. D. Griffin, "AI-enabled future crime," *Crime Science*, vol. 9, no. 1, pp. 1–13, 2020, doi: [10.1186/s40163-020-00123-8](https://doi.org/10.1186/s40163-020-00123-8).
- [2] M. S. Rana, M. N. Nobi, B. Murali, A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513. doi: [10.1109/ACCESS.2022.3154404](https://doi.org/10.1109/ACCESS.2022.3154404).
- [3] R. Chesney, D. Citron, "Deepfakes and the new disinformation war: The coming age of post-truth geopolitics," *Foreign Affairs*, vol. 98, p. 147, 2019. Available at: https://scholarship.law.bu.edu/shorter_works/76.
- [4] N. C. Köbis, B. Doležalová, I. Soraperra, "Fooled twice: People cannot detect deepfakes but think they can," *Iscience*, vol. 24, no. 11, p. 103364, doi: [10.1016/j.isci.2021.103364](https://doi.org/10.1016/j.isci.2021.103364).
- [5] M. Caldwell, J. T. Andrews, T. Tanay, L. D. Griffin, "AI-enabled future crime," *Crime Science*, vol. 9, no. 1, pp. 1–13, 2020, doi: [10.1186/s40163-020-00123-8](https://doi.org/10.1186/s40163-020-00123-8).
- [6] D. Milmo. (Feb 5, 2024). Company worker in Hong Kong pays out £20m in deepfake video call scam, The Guardian [Online]. Available: <https://www.theguardian.com/world/2024/feb/05/hong-kong-company-deepfake-video-conference-call-scam>. [Accessed: Aug. 05, 2024].
- [7] P. Kmieciak. (Dec. 21, 2023). Oszustwo 'na inwestycję w gazociąg'. Seniorka straciła pół miliona złotych, RMF24 [Online]. Available: https://www.rmf24.pl/regiony/wroclaw/news-oszustwo-na-inwestycje-w-gazociag-seniorka-stracila-pol-mili,Id.7223202#crp_state=1. [Accessed: Dec. 01, 2024].
- [8] M. Pawłowska. Oszustwo podczas inwestycji w akcje 'Baltic Pipe'. Policja Tomaszów Lubelski [Online]. Available: <https://tomaszow-lubelski.policja.gov.pl/itl/informacje/aktualnosci/141018,Oszustwo-podczas-inwestycji-w-akcje-Baltic-Pipe.html>. [Accessed: Dec. 01, 2024].
- [9] <https://policja.pl/pol/aktualnosci/242330,35-latka-stracila-blisko-150-000-zl-inwestujac-w-kryptowaluty.html>. [Accessed: Dec. 01 2024].
- [10] Y. Weng, S. S. Miryala, C. Khatri, R. Wang, H. Zheng, P. Molino, G. Tur, "Joint contextual modeling for ASR correction and language understanding," in *ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 4–8, 2020. New York, NY: IEEE, 2020, pp. 6349–6353. doi: [10.1109/ICASSP40776.2020](https://doi.org/10.1109/ICASSP40776.2020).
- [11] R. Cialdini, *Wwieranie Wpływu na Ludzi. Teoria i Praktyka (Influencing people: Theory and practice)*. Gdańsk: Gdańskie Wydawnictwo Psychologiczne, 1993.

Appendix 1. DEEP FRAME tool.

Module 1 – Technical analysis

Are there visible artifacts in the material?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Is the audio in sync with the video lip movement?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Is the area around the mouth and teeth more blurred or sharper than the rest of the image?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Does the voice sound natural?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Are there audio artifacts, such as clicks, distortions, or unnatural breaks?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Is there a shift in accent or tone, suggesting a foreign language influence?	<input type="checkbox"/> Yes <input type="checkbox"/> No
How do you rate the level of realism of the material?	<input type="checkbox"/> Low quality (easy to detect) <input type="checkbox"/> High quality (difficult to detect) <input type="checkbox"/> Hyper-realistic (virtually indistinguishable)
Are there signs of editing or manipulation (cuts, shifts)?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Are there background interferences, such as static noise, patterns, watermarks, halftones, or visual masks?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Does the content include logical errors or inconsistencies?	<input type="checkbox"/> Yes <input type="checkbox"/> No
Are there grammatical errors or incorrectly pronounced words in the content?	<input type="checkbox"/> Yes <input type="checkbox"/> No

Module 2 – Psychological analysis

The principle of social influence

The principle of reciprocity

The principle of commitment and consistency

The principle of social proof

The principle of authority

The principle of liking

The principle of scarcity

Examples	Frequency	Comments
----------	-----------	----------

Module 3 – Scope and context

Does the context of the publication suggest a specific target audience?	<input type="checkbox"/> Children <input type="checkbox"/> Eldery people <input type="checkbox"/> People with chronic diseases <input type="checkbox"/> Adutls <input type="checkbox"/> Men <input type="checkbox"/> Woman <input type="checkbox"/> Non-binary people <input type="checkbox"/> The voters <input type="checkbox"/> Believers of conspiracy theory <input type="checkbox"/> Others...
---	---

What is the purpose of the material?

- ☐ Parody
- ☐ Disinformation
- ☐ Financial fraud
- ☐ Discredit
- ☐ Blackmail
- ☐ Violation of privacy and dignity

Does the material use the image of a leader, expert, or celebrity?

- ☐ Yes
- ☐ No
- ☐ Provide the name and surname of the leader, expert, or celebrity

Did the material appear during the election campaign?

- ☐ Yes
- ☐ No

What is the risk level?

- ☐ Material may influence election results or political decisions
- ☐ Material may generate hate speech or escalate social conflicts
- ☐ Material may damage the reputation of a public or private person
- ☐ Material may lead to fraud
- ☐ Material may influence public opinion

Module 4 – Recommendations

Final steps

- ☐ Necessary confirmation (fact-checking) from trusted sources
 - ☐ Reporting to the platform administration or services
 - ☐ Educating recipients: publishing warnings and guides on manipulation
-

An Analysis of Cybersecurity Policy Compliance in Organisations

Hugues Hermann Okigui | Cape Peninsula University of Technology, South Africa | ORCID: 0009-0004-2221-5106

Johannes Christoffel Cronjé | Cape Peninsula University of Technology, South Africa | ORCID: 0000-0002-9838-4609

Errol Roland Francke | Cape Peninsula University of Technology, South Africa | ORCID 0000-0001-6029-9145

Abstract

In the contemporary digital landscape, cyberattacks and incidents have placed cybersecurity at the forefront of priorities in organisations. As organisations face cyber risks, it becomes imperative to implement and comply with various cybersecurity policies. However, due to factors such as policy complexity and resistance from employees, compliance can be a challenging task. The study, which took a comprehensive approach, investigated the variables that affect an organisation's adherence to cybersecurity policies. The findings of this study provide insights into the challenges and factors influencing compliance with cybersecurity policies in organisations. A case study design was chosen as part of a qualitative approach to answer the research question. For data gathering, semi-structured interviews were performed, and the existing documents were also considered when available to supplement interviews. The gathered data was meticulously organised, coded, and analysed using the Actor-Network Theory perspective, with a focus on its four moments of translation: problematisation, inter-essement, enrollment, and mobilisation. The analysis revealed that insider threats and phishing attempts are the two cyber threats

Received: 30.03.2024

Accepted: 02.08.2024

Published: 01.09.2024

Cite this article as:

H.H. Okigui, J.C. Cronjé, E.R. Francke "An analysis of cybersecurity policy compliance in organisations," ACIG, vol. 3, no. 2, 2024, pp. 303–321. DOI: 10.60097/ACIG/191942

Corresponding author:

Johannes Christoffel Cronjé, Cape Peninsula University of Technology, South Africa; E-mail: johannes.cronje@gmail.com

 0000-0002-9838-4609

Copyright:

Some rights reserved:
Publisher NASK



that affect organisations; behavioural challenges and enforcement limitations are the factors that influence and contribute to the non-compliance of cybersecurity policy; phishing exercises and policy development processes are used to enforce cybersecurity policies.

Keywords

cybersecurity policies, compliance challenges, insider threats, phishing attempts, Actor-Network Theory (ANT)

1. Introduction

Cybersecurity is not just a growing concern in specific regions but a global issue that affects countries around the world. This is evident in South Africa, where public and private organisations are constantly under threat from cyberattacks and incidents, leading to significant financial losses. The nation's high Internet access rate and increasing adoption of information and communication technology (ICT) have created a digital paradox situation where technological advances present countless opportunities for a country's development but also lead to a proliferation of cyber incidents and cyberattacks [1].

South Africa continues to be one of the most targeted nations in the world and Africa despite all the efforts [2–4]. The problem could be attributed to the fact that less focus has been put on human-related vulnerabilities, which represent the main target in most modern and recent cyberattacks and cyber incidents [5, 6].

This study aimed to analyse cybersecurity policy compliance in organisations. The study's results can be applied to direct and enforce agents' (end-users) compliance, through which cyber activities can be monitored and managed so as to minimise cyber incidents and cyberattacks within organisations. This study is underpinned by Actor-Network Theory (ANT), which is recognised as a social-technical theory. ANT is an increasingly used framework in social sciences, such as information systems, to examine the interactions between existing actors and how networks are built.

1.1. Aims, Objectives, and Research Questions

The aim of this study was to analyse the level of compliance with cybersecurity policies in organisations and to understand

the factors influencing this compliance. The objectives were as follows:

- To identify cyberattack and incidents registered by organisations.
- To understand factors that contribute to and influence non-compliance with cybersecurity policies in organisations.
- To examine how cybersecurity policy compliance is enforced in organisations.

The main question driving the study was: What are the factors influencing cybersecurity policy compliance in organisations? The outcome of this question could inform organisations on how to implement and enforce cybersecurity policies effectively, thereby improving their overall cybersecurity posture and reducing the risk of cyberattacks and incidents. The main question was refined with three sub-questions: (1) What are the cyberattacks and incidents that affect organisations? (2) What are the contributing and influencing factors to the non-compliance with cybersecurity policies in organisations? (3) How is cybersecurity policy compliance enforced in organisations?

2. Literature Survey

In keeping with the objectives and research questions, this literature survey covered cyberattacks and incidents, cybersecurity in organisations, and cybersecurity policy before briefly introducing and defending ANT as the lens through which analysis took place.

2.1. Cyberattack and Cyber Incidents

Millions of cyberattacks and incidents occur annually, causing significant financial losses and disruptions across various organisations [7]. As described by Hruza et al. [8], a cyberattack is an act perpetrated within cyberspace aimed at compromising cybersecurity objectives, including confidentiality, integrity, and availability, through activities such as data theft, modification, unauthorised access, destruction, or control of cyberspace infrastructure elements. Additionally, Ferreira [9] defines a cyber incident as a breach or imminent threat of breaching computer security policies, acceptable use policies, or standard security practices.

Organisations encounter diverse cyber threats due to evolving technologies and the constant development of new methods by malicious actors or hackers to compromise organisational assets' confidentiality, integrity, and authentication [10]. These threats

affect organisations, consumers, and stakeholders. Research indicates a shift in concern from traditional physical crimes to cybercrimes among organisations and their consumers [11].

Contemporary cybersecurity literature categorises cyber threats into four main groups: cyber terrorism, hacktivism, cybercrime, and cyber warfare [12]. Cybercrime, in particular, has escalated over the years, emerging as a significant concern for governments, private entities, and individuals [13]. Despite being a prevalent threat, cybercrime often receives less attention [12]. Reports highlight South Africa's vulnerability to cybercrimes, with statistics from Norton's cybercrime report in 2011 indicating high victim rates in South Africa and China [11, 14]. Furthermore, the Global Economic Crime and Fraud Survey for 2018 identified South Africa as the world's second most-targeted nation due to inadequate policing, underdeveloped laws, and inexperienced end-users [15]. Therefore, this study aims to identify the cyberattacks and incidents faced by organisations.

2.2. Cybersecurity in Organisations

The 2007 cyberattack on the Republic of Estonia thrust cybersecurity into prominence, showcasing the potential destabilisation of modern countries and organisations through ICT [16, 17]. Subsequently, cybersecurity has emerged as a significant concern for individuals and organisations globally, driven by the escalating frequency of cyberattacks and incidents, leading to substantial economic and safety repercussions for inadequately protected institutions [18].

Failure in cybersecurity not only results in costly losses for organisations but also poses critical risks to human lives, as hackers possess the capability to manipulate information systems, hindering the dissemination of evacuation alerts during emergencies [19]. The annual cost of cybercrime and economic espionage to the global economy is estimated to range from \$375 billion to \$575 billion [19], with South African organisations losing approximately 20 billion rand annually to cybercrimes [20].

Despite growing awareness of the importance of cybersecurity, some challenges persist in fostering global cooperation and alignment in combating cyber threats. A divergence in understanding cybersecurity among nations can affect collaborative efforts [21]. Nations, such as South Africa, have developed national cybersecurity strategies (NCSS) to articulate their understanding of

cybersecurity and to establish a harmonised framework of terminology and concepts [21].

Cybersecurity is defined as ‘the collection of tools, policies, security concepts, safeguards, risk management approaches, actions, training, best practices, assurance, and technologies to protect the cyber environment, organisation, and user assets’ [21]. South Africa’s vision regarding cybersecurity is to create a trusted and secure environment where ICT can be confidently utilised by individuals and organisations [22].

The South African perspective supports the importance of safeguarding human as well as non-human actors in cyberspace, thus aligning with Bada and Sasse’s [23] view that cybersecurity extends beyond protecting organisational assets to securing human users of ICT systems. Moreover, Mosca [24] asserts that effective cybersecurity measures enhance organisational sustainability and competitiveness by reducing vulnerability to cyber threats.

2.3. Policy and Compliance

Organisations have implemented technical measures to combat cybercrimes, including firewalls, Intrusion Detection Systems (IDS), Intrusion Prevention Systems (IPS), and other technological solutions [25]. However, relying solely on technology is insufficient against hackers’ evolving tactics, emphasising the necessity of integrating technical measures with robust security policies for effectiveness [10]. Security policies play a pivotal role in regulating and governing user behaviour within organisations, yet their implementation poses challenges [26].

Bayuk et al. [27] define policy as encompassing all regulations and laws aimed at maintaining organisational cybersecurity. A security policy outlines procedures and processes for employees to uphold the confidentiality, integrity, and availability of organisational resources. While having a cybersecurity policy is essential, experts stress the importance of compliance. Bulgurcu et al. [28] emphasise the need for organisations to understand and enhance employee compliance with existing policies to strengthen security measures.

Cavelty [29] asserts that cybersecurity policy is crucial for addressing global security challenges, focusing on common issues, such as vulnerability and privacy through regulatory frameworks. However, despite the presence of policies, there remains a gap between policy availability and employee practices, with non-compliance posing

significant risks [30, 31]. Consequently, organisations deploy awareness campaigns through various mediums like emails, posters, newsletters, and training modules [32].

Such approaches may only create a semblance of awareness, rather than fostering genuine compliance. Sannicolas-Rocca et al. [26] advocate for methods to improve and enforce employee adherence to security policies. Against this background, we examined the development, communication, and enforcement of cybersecurity policies within organisations.

2.4. Actor-Network Theory

Michael Callon, Bruno Latour, and John Law introduced ANT in the early 1980s where they emphasised the interactions and relationships within heterogeneous networks [33, 34]. ANT focuses on the construction, rather than the purpose of a network, with actors and networks being its core components. Both human and non-human entities are considered actors, and they are treated equally within networks. Actors can include technologies, tools, cultural meanings, and environmental conditions [35, 36].

In ANT, heterogeneity refers to networks comprising diverse elements [37, 38], where interactions among actors, such as people, technologies, texts, and others, form the basis of society [38]. Networks consist of established connections between actors, requiring movement and translation for their formation [39]. They facilitate collaboration among actors to address problems or create new entities [40].

Translation is a four-step process that involves problematisation, interessement, enrollment, and mobilisation. It is integral to network creation [41, 42] and involves persuading actors to accept roles and responsibilities that shape actor-network relationships [43].

Problematisation is the initiation phase of translation. Here, the focal actor identifies and describes the problem, aligns interests, and negotiates common goals [42, 44]. An obligatory passage point (OPP) represents a proposed solution during this phase [43]. Interessement follows problematisation and involves the recruitment of actors based on defined roles and responsibilities, persuading them of the problem's significance and proposed solution [42, 43]. Successful interessement leads to enrollment, where roles and responsibilities are assigned to recruited actors, and alliances and relationships are defined [43]. Enrollment succeeds

when actors accept their assigned roles, fostering a robust network of allies [42]. Upon completion of problematisation, interestment, and enrollment, mobilisation ensues, as designated spokespersons mobilise allies to act in alignment with their roles and responsibilities [42, 43].

3. Methods

This section discussed the case-study research approach and explain how the participating organisations and individual participants were sampled before discussing the collection of data through recorded Zoom interviews and the analysis of data through the lens of ANT. The analysis of the research was guided by ANT's four moments of translation – problematisation, interestment, enrollment, and mobilisation. These moments help in identifying the defined problem, negotiating interests, recruiting actors, assigning roles, and mobilising allies within the context of cybersecurity policy compliance.

3.1. Research Approach and Sampling

This study focused exclusively on exploring organisation employees' attitude towards organisation cybersecurity policy, employing a case study design for its ability to investigate phenomena within their natural settings [23]. The flexibility of case study design allows for the examination of various research questions while considering contextual influences [23]. The qualitative case study methodology is deemed valuable for studying complex phenomena within their contexts [23].

Despite challenges in obtaining sufficient samples due to the topic's sensitivity, three South African-based organisations were included in the study, with one functioning as a cybersecurity service provider [23]. The selection criteria focused on organisations with cybersecurity departments responsible for maintaining the Confidentiality, Integrity, Authenticity (CIA) triad [23]. The use of pseudonyms ensured anonymity for the organisations: HollanRaph for Case 1, NoahGabi for Case 2, and LenJo for Case 3.

HollanRaph, a large higher education institution with over 5000 staff located in Gauteng province, demonstrates a strong commitment to cybersecurity through the establishment of a dedicated cybersecurity department and policy. This organisation functions as a dynamic network involving both human actors and non-human actants. NoahGabi, another large higher education institution in

Western Cape province, boasts over 32,000 students and 5000+ staff, positioning itself as one of the largest institutions in the region. Despite its size, NoahGabi acknowledges the importance of cybersecurity and maintains a dedicated team onsite to address related concerns. Lastly, LenJo, a small IT services and consulting firm based in Gauteng province, plays a significant role in the South African information technology (IT) landscape. Specialising in business-to-business (B2B) ICT solutions, LenJo serves a diverse clientele ranging from small businesses to multinational enterprises. With a focus on business process digitalisation, cybersecurity services, and ICT skills development, LenJo aims to emerge as a leader in its field.

Four participants were purposively selected from these organisations and interviewed via Zoom, with the interviews being recorded. Participant 1 holds the position of manager: IT risk and compliance with over 10 years of experience, contributing to the HollanRaph case. Participant 2, a senior systems engineer specialising in networks and information security with over 10 years of experience, also pertains to the HollanRaph case. Participant 3, the chief executive officer (CEO) and security specialist at LenJo, brings over 9 years of experience to the study. Participant 4, serving as manager of IT strategic services, has over 10 years of experience and is associated with the NoahGabi case.

3.2. Data Analysis

Interview data were transcribed, cleaned, and analysed. The analysis aimed to extract meaningful information from the collected data through transcription, facilitating easier management and analysis [45]. Employing ANT's four moments of translation – problematisation, interessement, enrollment, and mobilisation – guided the analysis from three perspectives: the existence of actors (human and non-human), creation of networks, and interactions and relationships [42, 43]. These moments were utilised to identify the defined problem, negotiate interests, recruit actors, assign roles, and mobilise allies [42, 43]. ANT proved beneficial in identifying actors, including focal actors, and examining network creation and actor relationships, enhancing the understanding of the phenomenon [43]. By considering both human and non-human entities, ANT allowed us to obtain insights into how connections and interactions contributed to network formation, which was particularly relevant in understanding cybersecurity policy involving various entities [39].

4. Results and Discussion

Using ANT as a lens, our analysis focused on actors, networks, and moments of translation. We identified the actors involved in cybersecurity activities, examined their roles, and assessed the implications. Similarly, the study explored the networks existing within cybersecurity activities in South African organisations. The moments of translation, involving negotiation among actors within heterogeneous networks, helped to understand the complex and multidimensional nature of cybersecurity activities, as described by Dlamini and Modise [14].

4.1. Actors

In ANT, actors encompass both human and non-human entities capable of influencing their environment [46]. Both humans and non-humans are integral to cybersecurity activities. Human actors, including technical (IS/IT personnel) and non-technical counterparts, play roles delegated or voluntarily assumed within organisations involved in cybersecurity. Technical personnel have various roles, such as IT risk and compliance managers, IT strategic services managers, security specialists, and systems engineers. At the same time, non-technical actors include business personnel, end-users, managers, clients, and partners. Non-human actors directly or indirectly involved in cybersecurity activities include cybersecurity policies, phishing exercises, computer systems and networks, and security awareness programmes. These components encompass written policies, phishing simulations, computer systems and networks, and security awareness initiatives aimed at informing and educating organisational personnel about potential threats and best practices [46].

4.2. Networks

In cybersecurity policy compliance, actor networks facilitate collaborative problem-solving and entity creation [40]. Networks, heterogeneous in nature, comprise diverse actors, both human and non-human, with an actor potentially belonging to multiple networks. Major actor-networks in this context include the organisation, risk committee, IT managers, business managers, technologists, and end-users. Each network has distinct roles and responsibilities in managing cybersecurity policies [40]. The executive committee, comprising leadership personnel, drafts and enforces cybersecurity policies and standards. Business managers oversee compliance with policies and processes to achieve

organisational objectives. Technologists, including IT engineers and security specialists, develop training and methods for cybersecurity activities. Internal end-users utilise organisational information systems, while external end-users, such as clients and partners, also face cybersecurity risks [40].

4.3. Moments of Translation

In ANT, translation is concerned with negotiations that occur within networks. The negotiations are shaped by the interactions that happen among actors, which are influenced by various interests. Transformations are observed within organisations based on negotiations and activities. There are four moments in the process of translation: problematisation, interessement, enrollment, and mobilisation [47].

Problematisation: As described by Jessen and Jessen [43], this is where the focal actor(s) identify and define the problem. In the context of ANT, a problem is not necessarily a broken thing but requires a solution, in some cases, an improvement [48]. Organisations are challenged with cyberattacks and incidents particularly with insider threats and phishing attempts type. The insider threats and phishing attempts are from different sources. Some of the sources are internal, and others are external. The internal sources are related to the end-users' behaviours and are either conscious or unconscious. Irrespective of the consciousness or the unconsciousness of end-users' behaviours, cyberattacks and incidents such as phishing attacks and insider threats are occasioned.

Insider threats and phishing attempts represent a significant cybersecurity problem for organisations. Thus, effective measures are needed to address the problem. Another existing problem is behavioural challenges. As stated by a participant, despite several awareness materials put in place by organisations, it is still difficult to instigate a change of mind among end-users. According to another participant, the lack of compliance with existing cybersecurity policies poses a critical problem:

So, the current attack we experience mostly is around phishing. We get a significant amount of phishing attempts. Directed to staff and directed to students. That dominates our cybersecurity awareness efficiency because if I look at the incidents we experienced over the past years, 90% of those would be phishing-related cyber incidents. (L49-54_P1_NoahGabi)

The challenges are behavioural challenges. It's just a change in mindset because we share quite a few awareness materials. So, on quite a few platforms, we still have end-users who would fall for a phishing attempt, you know? Given the kinds of initiatives that we're trying to put in place, you would expect that there would be quite a bit of improvement in behaviour. That's one of the challenges. (L220-227_P2_HollanRaph)

Interessement: The Interessement phase starts from the moment a problem is identified. At this phase, the links between the interests of different actors and allies are aligned and strengthened [47]. The alignment of actors' interests is done through negotiations. The negotiations are based on each actor's interests and the roles they may play in the network. To do so, focal actor(s) explain to others and allies how their own goals can be achieved by joining the network. As described by Iyamu and Mgudlwa [48], this phase is important because the alignment of different interests can contribute to addressing what was problematised. Additionally, the interests are various and can be expressed in different ways. Some people's interests can be based on their obligations, positions, or/ or duties in an organisation. For others, the interests can be based on their business goals, passions, or the implications that cybersecurity policy or cyberattacks and incidents could have on them.

Some organisations are facing difficulties in enforcing their cybersecurity policies. As emphasised by a participant, this is due to the nature of the environment in some organisations, particularly those having multiple natures of end-users in their environment. Unlike sectors, such as healthcare and banking, the educational sector faces challenges in enforcing its cybersecurity policies. Using the one-size-fits-all method for awareness programmes or materials has not been working. So, there is a need for a different approach that could accommodate various natures of end-users. In this context, failing to tailor an awareness approach to all end-users is a focal point of interest for cybersecurity makers:

I've worked in many different organisations, and when you take a banking environment where it's very regulated, right? Or a mining, one of the mining organisations, it's enforced in terms of compliance awareness exercises. If you don't do the training, there's repercussions for that. You don't, you're locked out of your computer. But it's a different environment, and we are unable to enforce those kinds of hard and first rules to say we'll lock you out

because we're working with students and lecturers. So, business needs to continue. So, it's a bit of a balancing act. (L229-237_P1_ HollanRaph)

Compliance is always a challenge. The fact that we are a higher education structure means that we encourage that idea of openness for collaboration, and the difficulty is that it creates complexity and challenges because we are not dealing with one state of staff. We are dealing with many different types of staff, such as academics, students, and many others, and I think that is the challenge. The challenge is tailoring a program that suits everyone. So, you need to engage with people on a regular basis, so I think there is difficulty in compliance with that because you get to deal with such a broad circle of people. I think that is the challenge that we are looking into and actively trying to address. (L88-98_P1_ NoahGabi)

Enrollment: It is a critical phase in the process of translation. In this phase, actors are brought together in the same network with the common purpose of finding an effective measure to address the identified problem. It is also about developing alliances and investigating how the actors align in the common objective of developing an effective cybersecurity policy and awareness programmes. To enforce, educate, and inform end-users with the most important aim to enforce. Furthermore, the existence of cybersecurity policy and awareness programmes, such as simulated phishing emails, indicate enrollment and organisation with the objective of addressing the problem. Another point is to motivate those who do not really understand the criticality behind the whole intention of securing the systems. A participant highlighted that the reluctance of those actors is based on the approach used when communicating with them. The participant continued saying that they sometimes have to get involved in politics to stimulate them:

Well, I'm the risk and compliance manager in ICT. I look after governance, so ICT policies, frameworks, standards, processes, and procedures. (L51-53_P1_ HollanRaph)

I'll give you an example: you walk to a person and say, listen, I need to check that your antivirus endpoint firewall is up and running. They are not going to like it because they are busy, but when you say listen, If I don't do this when you are doing your own personal online banking, people are going to be able to see your credentials and take your

money. Then suddenly it changed because it's no longer. I think you're wasting my time, but it's about their money or their well-being. (L410-416_P1_ LenJo)

Mobilisation: It is the last phase, and it takes place when the problematisation, interessement, and enrollment phases are completed [42]. This phase is important because it is where the main actor makes sure that others behave with respect to their assigned roles and responsibilities [43]. The mobilisation phase also aims to mobilise developed networks and maintain proposed solutions to address identified problems effectively. The purpose of mobilisation was to keep other end-users focused and conscious about the issues of cyber threats, in particular phishing attempts and insider threats. This was done through the organisation's cybersecurity policies and activities like phishing exercises conducted quarterly. Phishing exercises were used to evaluate the level of compliance or vigilance of actors such as end-users. This also helped to assess their capability of detecting potential cyber threats. Then, collected outcomes could be an important resource as they highlighted gaps and pointed out where more attention was needed. Once the gaps are identified, improvements could be made in cybersecurity policies and materials that create awareness:

It is through fishing exercises. So, they have been quarterly, and we do get reports on them that tell us how many people clicked on the link. It would tell us who, specifically, which department and what information they divulged. So that gives us an indication. Then, we're able to target specific training for those individuals per area. (L245-250_P2_ HollanRaph)

5. Conclusions

This section provides a summary of the key findings, an answer to the research questions, a discussion of this study's limitations, and recommendations for further research.

5.1. Summary of Findings

The following is a summary of the findings obtained from the processed qualitative analysis:

- Behavioural challenges refer to end-users' attitudes and behaviours towards cybersecurity measures initiated by organisations, including resistance to complying with cybersecurity

policies and a tendency to fall for phishing attempts. Enforcement limitations involve a lack of suitable cybersecurity policies and awareness programmes that align with the specific needs of the organisation's end-users. Insider threats encompass both conscious and unconscious cyber risks generated by personnel within the organisation. Phishing attempts are fraudulent efforts to steal sensitive information, such as login credentials, often delivered via email or SMS. Phishing exercises simulate real-world phishing attacks to test the readiness of staff or end-users in identifying cyber threats and to evaluate the effectiveness of existing awareness programmes. The policy development process involves creating a cybersecurity policy that considers all phases—drafting, review, and approval—and requires collaboration with relevant stakeholders to ensure it meets the organisation's unique context.

5.2. Answers to Research Questions

Research sub-question 1: What are the cyberattacks and incidents that affect organisations?

The analysis conducted in Section 4.3 showed that organisations are particularly challenged with the following:

- *Insider threats*: The analysis also revealed that insider threats involved staff or internal end-users with authorised access, and their occurrence was either conscious or unconscious.
- *Phishing attempts*: On the other hand, phishing attempts, usually in the form of email or SMS, were fraudulent attempts perpetrated by external individuals with the intention of stealing sensitive information, such as end-users or staff login credentials.

Research sub-question 2: What are the factors that influence and contribute to non-compliance with cybersecurity policies in organisations?

The analysis showed that the factors influencing and contributing to non-compliance with the organisation's cybersecurity policies are as follows:

- *Behavioural challenges*: The behavioural challenges concern internal end-user mindsets and attitudes towards proposed cybersecurity policies. Despite awareness initiatives taken by organisations, internal end-users were not adhering to the security measures available to them.

- *Enforcement limitations:* The enforcement of limitations is the fact that some organisations are failing to develop adequate and balanced cybersecurity policies to meet their heterogeneous environment contexts. Proposed policies are sometimes not suitable for the business sector they are in. Consequently, not all internal end-users can be targeted. For example, higher education and banking-type environments cannot consider similar aspects when developing cybersecurity policies and awareness programmes. Some organisations cannot have a one-size-fit cybersecurity policy.

Research sub-question 3: How is cybersecurity policy compliance enforced in organisations?

According to the analysis provided in Section 4.3, organisations enforce their cybersecurity policy compliance using the following:

- *Phishing exercises:* The analysis revealed that periodically, phishing exercises, such as simulated phishing emails, were initiated. The main purpose of this approach is to evaluate the readiness of internal end-users or staff to see if they are well equipped and capable of identifying and avoiding falling into some types of cyber threats. Furthermore, phishing exercise reports could indicate where improvement is needed in the current proposed solutions.
- *Policy development process:* The analysis showed that the cybersecurity policy development process should follow a collaborative and inclusive approach, with participation of organisation stakeholders. Potential policies should be drafted first, reviewed, and then submitted for approval.

5.3. Contribution of the Research

Theoretical contributions: The study contributes to the academic literature, especially the fact that very little has been done in the area of cybersecurity studies through the ANT concept, especially using the four moments of translation. ANT is employed to explore the actors and networks involved in cybersecurity activities within organisations. It helps in understanding the roles of human and non-human entities in cybersecurity, such as IT personnel, business managers, end-users, clients, partners, cybersecurity policies, phishing exercises, computer systems, networks, and security awareness programmes.

Practical contributions: This study is important, as we hope the result will continuously assist organisations with their cybersecurity policy challenges and the persistently growing number of cyberattacks

and incidents. The findings could help us better understand these challenges and develop more contextualised cybersecurity policies to fit organisational environments.

5.4. Limitations of the Study

Due to the sensitivity of the topic, some organisations were reluctant to participate in the study. Thus, this study was limited in terms of participants. The researcher emphasises the concept of caution transferability of findings. The researchers suggest that the results of this study should be applicable to organisations with similar settings.

5.5. Recommendations for Further Research

The analysis presented in this study reveals that one of the challenges faced by organisations is enforcement limitations. This means that some organisations do not have the capacity or fail to develop cybersecurity policies that are suitable for their environment. Based on this, it would be interesting to select two different sectors and then conduct a comparative analysis.

References

- [1] S. Mabunda, "Cybersecurity in South Africa: Towards Best Practices," in *CyberBRICS: Cybersecurity Regulations in the BRICS Countries*. Cham: Springer International Publishing, 2021, pp. 227-270, doi: [10.1007/978-3-030-56405-6_6](https://doi.org/10.1007/978-3-030-56405-6_6).
- [2] N. Kshetri, "Cybercrime and cybersecurity in Africa," *Journal of Global Information Technology Management*, vol. 22, no. 2, pp. 77-81, 2019, doi: [10.1080/1097198X.2019.1603527](https://doi.org/10.1080/1097198X.2019.1603527).
- [3] M. Evans, L.A. Maglaras, Y. He, H. Janicke, "Human behaviour as an aspect of cybersecurity assurance," *Security and Communication Networks*, vol. 9, no. 17, pp. 4667-4679, 2016, doi: [10.1002/sec.1657](https://doi.org/10.1002/sec.1657).
- [4] N. Kortjan, R. Von Solms, "A conceptual framework for cybersecurity awareness and education in SA," *South African Computer Journal*, vol. 52, no. 1, pp. 29-41, 2014, doi: [10.18489/sacj.v52i0.201](https://doi.org/10.18489/sacj.v52i0.201).
- [5] Gundu, T. (2019). "Acknowledging and Reducing the Knowing and Doing gap in Employee Cybersecurity Compliance," *Proceedings of the 14th International Conference on Cyber Warfare and Security*, N. van der Waag-Cowling, L. Leenen, Eds. Stellenbosch University, South Africa, February 28-March 1, 2019.
- [6] J. Abawajy, "User preference of cybersecurity awareness delivery methods," *Behaviour & Information Technology*, vol. 33, no. 3, pp. 237-248, 2014, doi: [10.1080/0144929X.2012.708787](https://doi.org/10.1080/0144929X.2012.708787).
- [7] J.S. Nye Jr, "Deterrence and dissuasion in cyberspace," *International Security*, vol. 41, no. 3, pp. 44-71, 2017, doi: [10.1162/ISEC_a_00266](https://doi.org/10.1162/ISEC_a_00266).

- [8] P. Hruza, R. Sousek, S. Szabo. (2014). "Cyberattacks and attack protection," in *World multi-conference on systemics*, vol. 18, pp. 170–174 [Online]. Available: https://www.iis.org/CDs2014/CD2014SCI/SCI_2014/PapersPdf/SA975KW.pdf [Accessed: Jul. 21, 2018].
- [9] F.W. Ferreira. (2012). "*NIST publishes computer security incident handling guide*" [Online]. Available: <https://www.hlregulation.com/2012/08/16/nist-publishes-computer-security-incident-handling-guide/> [Accessed: Jun. 22, 2018].
- [10] N.S. Safa et al., "Information security conscious care behaviour formation in organisations," *Computers & Security*, vol. 53, pp. 65–78, 2015, doi: [10.1016/j.cose.2015.05.012](https://doi.org/10.1016/j.cose.2015.05.012).
- [11] N. Kshetri, *Cybercrime and cybersecurity in the global south*. London: Palgrave Macmillan, 2013, doi: [10.1057/9781137021946](https://doi.org/10.1057/9781137021946).
- [12] K. Quigley, C. Burns, K. Stallard, "Cyber Gurus: A rhetorical analysis of the language of cybersecurity specialists and the implications for security policy and critical infrastructure protection," *Government Information Quarterly*, vol. 32, no. 2, pp. 108–117, 2015, doi: [10.1016/j.giq.2015.02.001](https://doi.org/10.1016/j.giq.2015.02.001).
- [13] D. Marsh. (2017). "*Are ethical hackers the best solution for combating the growing world of cybercrime?*" Unpublished doctoral dissertation, University Honors College, Middle Tennessee State University, TN [Online]. Available: <https://jewelscholar.mtsu.edu/bitstreams/2a3a0af1-5bdf-41dc-859a-8c44c92ffcfb/download> [Accessed: Jul. 21, 2018].
- [14] Dlamini, Z., Modise, M., 2012. Cyber security awareness initiatives in South Africa: a synergy approach. In 7th International Conference on Information Warfare and Security. pp. 1–10. [Online] Available at: <http://hdl.handle.net/10204/5941> [Accessed: Jan. 21, 2020].
- [15] Citizen. (2018). *Presidency website back up after hack*. [Online]. Available: <https://citizen.co.za/news/south-africa/1972813/presidency-website-back-up-after-hack/> [Accessed: Apr. 4, 2019].
- [16] N. Gcaza, R. von Solms, J.J. van Vuuren. (2015). "An ontology for a national cybersecurity culture environment," in: *HAISA*, pp. 1–10. [Online]. Available: <https://www.researchgate.net/publication/306292545> [Accessed: Apr. 16, 2019].
- [17] A. Kozłowski, "Comparative analysis of cyberattacks on Estonia, Georgia and Kyrgyzstan," *European Scientific Journal (ESJ)*, vol. 3, pp. 237–245, 2014, <https://www.researchgate.net/publication/260107032>.
- [18] N. Gcaza, R. Von Solms, "A strategy for a cybersecurity culture: A South African perspective," *The Electronic Journal of Information Systems in Developing Countries*, vol. 80, no. 1, pp. 1–17, 2017, doi: [10.1002/j.1681-4835.2017.tb00590.x](https://doi.org/10.1002/j.1681-4835.2017.tb00590.x).
- [19] J.P. Kesan, C.M. Hayes, "Creating a 'Circle of Trust' to Further Digital Privacy and Cybersecurity Goals," *Forthcoming, Michigan State Law Review, Illinois Public Law Research Paper No. 13-03, Illinois Program in Law, Behavior and Social Science Paper No. LBSS13-04*, pp 1475-1559, 2014, doi: [10.2139/ssrn.2135618](https://doi.org/10.2139/ssrn.2135618).
- [20] N. Kshetri, "Cybercrime and cybersecurity issues in the BRICS economies," *Journal of Global Information Technology Management*, vol. 18, no. 4, pp. 245–249, 2015, doi: [10.1080/1097198X.2015.1108093](https://doi.org/10.1080/1097198X.2015.1108093).

- [21] E. Luijff, K. Besseling, P. De Graaf, "Nineteen national cybersecurity strategies," *International Journal of Critical Infrastructures* 6, vol. 9, no. 1–2, pp. 3–31, 2013, doi: [10.1504/IJCIS.2013.051608](https://doi.org/10.1504/IJCIS.2013.051608).
- [22] R. Von Solms, J. Van Niekerk, "From information security to cybersecurity," *Computers & security*, vol. 38, pp. 97–102, 2013, doi: [10.1016/j.cose.2013.04.004](https://doi.org/10.1016/j.cose.2013.04.004).
- [23] M. Bada, A. Sasse, "Cybersecurity awareness campaigns: Why do they fail to change behaviour?" Oxford: Global Cyber Security Capacity Centre, University of Oxford, 2014, arXiv:1901.02672.
- [24] M. Mosca. (2015). "Cybersecurity in an era with quantum computers: will we be ready?" IACR Cryptology ePrint Archive, p. 1075, [Online]. Available: <https://eprint.iacr.org/2015/1075.pdf> [Accessed: Apr 19, 2019].
- [25] H.M. Said et al., "An integrated approach towards a penetration testing for cyberspaces," *European Journal of Computer Science and Information Technology*, vol. 3, no. 1, pp. 108–128, 2015, doi: [10.1186/s12913-018-3161-3](https://doi.org/10.1186/s12913-018-3161-3).
- [26] T. Sannicolas-Rocca, B. Schooley, J.L. Spears, "Designing effective knowledge transfer practices to improve IS security awareness and compliance," in *Proceedings of the 2014 47th Hawaii International Conference on System Sciences*, pp. 3432–3441, IEEE, doi: [10.1109/HICSS.2014.427](https://doi.org/10.1109/HICSS.2014.427).
- [27] J.L. Bayuk et al., *Cybersecurity policy guidebook*. Hoboken, NJ: John Wiley, 2012, doi: [10.1002/9781118241530](https://doi.org/10.1002/9781118241530).
- [28] B. Bulgurcu, H. Cavusoglu, I. Benbasat, "Information security policy compliance: An empirical study of rationality-based beliefs and information security awareness," *MIS Quarterly*, vol. 34, no. 3, pp. 523–548, 2010, doi: [10.2307/25750690](https://doi.org/10.2307/25750690).
- [29] M.D. Cavelti, "Breaking the cybersecurity dilemma: Aligning security needs and removing vulnerabilities," *Science and Engineering Ethics*, vol. 20, no. 3, pp. 701–715, 2014, doi: [10.1007/s11948-014-9551-y](https://doi.org/10.1007/s11948-014-9551-y).
- [30] D.C. Streeter, "The effect of human error on modern security breaches," *Strategic Informer: Student Publication of the Strategic Intelligence Society*, vol. 1, no. 3, p. 2, 2013, <https://www.researchgate.net/publication/260107032>.
- [31] J. Blythe. (2013). "Cybersecurity in the workplace: Understanding and promoting behaviour change," in: *Proceedings of CHItaly 2013 doctoral consortium*, pp. 92–101 [Online]. Available: <https://nrl.northumbria.ac.uk/id/eprint/14720> [Accessed: Apr. 16, 2019].
- [32] E. Albrechtsen, J. Hovden, "Improving information security awareness and behaviour through dialogue, participation and collective reflection. An intervention study," *Computers & Security*, vol. 29, no. 4, pp. 432–445, 2010, doi: [10.1016/j.cose.2009.12.005](https://doi.org/10.1016/j.cose.2009.12.005).
- [33] B. Czarniawska, "Actor-network theory," in: *The Sage handbook of process organisation studies*, A. Langley, H. Tsoukas, Eds. Los Angeles, CA: Sage, 2016, pp. 160–173, doi: [10.4135/9781473957954.n10](https://doi.org/10.4135/9781473957954.n10).
- [34] B.J. Greenhough, "Actor-network theory," in *International encyclopedia of geography: People, the earth, environment and technology*, Wiley Online Library, 2016, pp. 1–7, doi: [10.1002/9781118786352.wbieg0532](https://doi.org/10.1002/9781118786352.wbieg0532).

- [35] O. Hanseth, M. Aanestad, M. Berg, "Guest editors' introduction: Actor-network theory and information systems. What's so special?," *Information Technology & People*, vol. 17, no. 2, pp. 116–123, 2004, doi: [10.1108/09593840410542466](https://doi.org/10.1108/09593840410542466).
- [36] J. Scott, *Social network analysis*. London: Sage, 2017, doi: [10.4135/9781529716597](https://doi.org/10.4135/9781529716597).
- [37] Y. Shim, D.H. Shin, "Analysing China's fintech industry from the perspective of actor-network theory," *Telecommunications Policy*, vol. 40, no. 2, pp. 168–181, 2016, doi: [10.1016/j.telpol.2015.11.005](https://doi.org/10.1016/j.telpol.2015.11.005).
- [38] B. Latour, *Reassembling the social: An introduction to actor-network-theory*. Oxford: Oxford University Press, 2007.
- [39] R. Dankert. (2010). *Using actor-network theory (ANT) doing research*. [Online]. Available: www.ritskedankert.nl/publications [Accessed: Apr. 14, 2019].
- [40] T. Iyamu, T. Sekgweleo, "Information systems and actor-network theory analysis," *International Journal of Actor-Network Theory and Technological Innovation (IJANTTI)*, vol. 5, no. 3, pp. 1–11, 2013, doi: [10.4018/jantti.2013070101](https://doi.org/10.4018/jantti.2013070101).
- [41] I. Williams, *The role of community based networks in the development of rural broadband. The case of Djurslandsnet in Denmark and lessons for rural sub-Saharan Africa*. Munich: Grin Publishing, 2014.
- [42] C. Costa, P. Cunha, "The social dimension of business models: An actor-network theory perspective," in 21st Americas Conference on Information Systems, AMCIS 2015, Puerto Rico, August 13-15, 2015. Association for Information Systems, 2015. [Online]. Available: <https://aisel.aisnet.org/amcis2015/e-Biz/GeneralPresentations/25> [Accessed: Nov. 14, 2023].
- [43] Jessen, J. D., Jessen, C. "Games as Actors - Interaction, Play, Design, and Actor Network Theory," *International Journal on Advances in Intelligent Systems*, vol. 7, no. 3–4, pp. 412–422, 2014.
- [44] R. Heeks, C. Stanforth, "Technological change in developing countries: Opening the black box of process using actor-network theory," *Development Studies Research*, vol. 2, no. 1, pp. 33–50, 2015, doi: [10.1080/21665095.2015.1026610](https://doi.org/10.1080/21665095.2015.1026610).
- [45] S.R. Ponelis, "Using interpretive qualitative case studies for exploratory research in doctoral studies: A case of information systems research in small and medium enterprises," *International Journal of Doctoral Studies*, vol. 10, no. 1, pp. 535–550, 2015, doi: [10.28945/2339](https://doi.org/10.28945/2339).
- [46] S. Edwards, *Doing actor-network theory: Integrating network analysis with empirical philosophy in the study of research into genetically modified organisms in New Zealand*. Doctoral dissertation, Lincoln University, Lincoln, UK, 2014. [Online]. Available: <https://hdl.handle.net/10182/6744> [Accessed: Apr. 14, 2019].
- [47] A. Wæraas, J.A. Nielsen, "Translation theory 'translated': Three perspectives on translation in organisational research," *International Journal of Management Reviews*, vol. 18, no. 3, pp. 236–270, 2016, doi: [10.1111/ijmr.12092](https://doi.org/10.1111/ijmr.12092).
- [48] T. Iyamu, S. Mgudlwa, "Transformation of healthcare big data through the lens of actor network theory," *International Journal of Healthcare Management*, vol. 11, no. 3, pp. 182–192, 2018, doi: [10.1080/20479700.2017.1397340](https://doi.org/10.1080/20479700.2017.1397340).

Jobs Exposure to Generative AI: Ongoing Study by NASK-PIB and ILO

Marek Troszyński | NASK-PIB Office for Analysis and Research, Collegium
Civitas, Warsaw | ORCID: 0000-0002-3653-4018

Abstract

The rapid advancement of artificial intelligence (AI), including generative AI (GenAI), raises important questions about its impact on the labour market and employment structure. This study examines the extent to which various occupations are exposed to GenAI by developing an index to identify potential shifts in the nature of work. The analysis focuses on specific occupational tasks that may be affected to varying degrees by the proliferation of AI tools. The study categorises occupations into four groups: susceptible to automation (Automation potential), subject to augmentation by GenAI (Augmentation potential), characterised by significant uncertainty (Big unknown), and not susceptible to technological change (Not affected). The research was conducted in three stages: assessing occupational exposure, verifying findings with expert analysis, and extrapolating results to 30,000 tasks across 2,500 occupations, with the support of ChatGPT-4. The findings enable estimates of the occupational groups most “at risk” from GenAI and contribute to macroeconomic forecasts for the Polish labour market.

Keywords

genAI, automation, new technologies, labour market, occupation index

Received: 06.12.2024

Accepted: 16.12.2024

Published: 31.12.2024

Cite this article as:

Troszyński, M, “Jobs exposure to generative AI: Ongoing study by NASK-PIB and ILO,” ACIG, vol. 3, no. 2, 2024, 322–327. DOI: 10.60097/ACIG/201152

Corresponding author:

Marek Troszyński, NASK-PIB Office for Analysis and Research, Collegium Civitas, Warsaw. E-mail: marek.troszynski@nask.pl

 0000-0002-3653-4018

Copyright:

Some rights reserved
(CC-BY):

Marek Troszyński
Publisher NASK



1. Introduction

The debate about the capabilities and limitations of artificial intelligence (AI) has dominated public space in the recent months. Much of this discussion pertains to economic issues – changes which the introduction of AI tools brings about for the global economy. However, an emotional and controversial question is whether AI tools will be able to replace humans in the labour market.

Translating this into the language of social studies, we talk about the fear of automation and technological unemployment. This problem is analysed, among others, in the paper titled ‘Who’s afraid of automation? Examining determinants of fear of automation in six European countries’ [1]. Its authors used data from the Central European Social Survey conducted at the turn of 2021 and 2022 in six Central European countries. Analysis on a sample of 6600 economically active people showed that one in six respondents was afraid of the impact of automation. More importantly, the more an economic sector is saturated with technology, the greater the fear of automation. Therefore, both knowledge of technology and its presence in one’s workplace exacerbate the fear of automation [1].

Workplace automation (robotisation) historically preceded the current process involving the popularisation of AI tools (in particular those based on generative artificial intelligence (GenAI)) that may replace human tasks. Therefore, it is ever so pertinent to ask the following questions: What jobs are most exposed to this process? How can we study that?

This problem is solved, thanks to macroeconomic forecasts estimating the impact of GenAI on economy and the labour market. This not only allows large economic organisations to plan efforts and determine long-term strategies but also equips labour market actors with the knowledge needed to decide how to shape and design their career paths.

This precisely – developing an index of occupations which would allow for estimating their exposure to GenAI tools – is what we have set out to do within the framework of the project titled, ‘The potential impact of generative artificial intelligence on job quantity and quality in Poland’, implemented, as commissioned by the Ministry of Digital Affairs, at the Research and Academic Computer Network (NASK) in cooperation with International Labour Organisation (ILO). The detailed assumptions for this project are described in the inception report [2].

2. The Impact of GenAI on the Labour Market – A Methodological Review

In the recent years, the issue of developing an index for the possible use of GenAI in respective occupations has been described in literature for many times. There are a few approaches that are worth mentioning. The most popular index is the one developed by Felten et al. [3,4] – AI occupational exposure (AIOE). It measures job exposure to AI, enabling the assessment of the degree to which various occupations are exposed to AI impacts, without determining whether these impacts are positive or negative. In this approach, the researchers invoked 10 AI applications specified by the Electronic Frontier Foundation, such as abstract strategy games, real-time video games, image recognition, visual question answering, image generation, reading comprehension, language modelling, translation, speech recognition, and instrumental track recognition. These AI applications were collated together with 52 human abilities (such as oral comprehension, oral expression, inductive reasoning, arm-hand steadiness, etc.) collected in the Occupational Information Network (O*NET) database developed by the US Department of Labor. Each of over 800 occupations is perceived as a weighted combination of 52 human abilities. Felten's team sent their study questionnaire to gig workers at Amazon Mechanical Turk (mTurk) and collected 1800 responses [4]. The respondents had to assess whether a specific task could be performed by GenAI tools. Exposure at skill level was calculated as the total of connections between AI applications and human skills. Then, AIOE was calculated for each occupation (consisting of specific skills), accounting for how important and widespread these skills are in a given occupation [4, pp. 3–4].

In their modified approach published in 2023, the authors singled out 'language modelling' as the key skill to be replaced by GenAI. They then specified as to what extent this skill is important for the respective occupation.

Summing up this approach, the authors specified occupations that are most vulnerable to automation. These included telemarketers, English language and literature teachers, foreign language and literature teachers, history teachers, clinical psychologists, advisors, and local government workers [4, p. 14]. Viewing the data from the industry perspective, the most vulnerable are legal services, financial services (trading in securities), insurance and employee benefit funds as well as universities and training institutions [4, p. 15].

In Poland, the Polish Economic Institute published in 2024 its report titled 'AI na polskim rynku pracy' ('AI on the Polish labour market') [5], presenting estimates concerning the Polish labour market based on the aforementioned AIOE index. The researchers estimated that there are 3.68 million Poles in the 20 occupations most exposed to AI [5, p. 24].

Another approach involves analysing the demand for certain skills in a given labour market. For example, Acemoğlu et al. [6] analysed online job postings and their specific skill requirements. They considered such sources as burning glass technologies and job search sites. In the research process, it was necessary to identify the skills and technologies advertised. The weakness of this method lay in the skew towards jobs posted online. Therefore, this approach did not work in countries where most job vacancies were advertised offline.

Yet another method found in source literature is experiment-based analysis. In Peng et al. [7], researchers performed a controlled experiment among professional programmers who were given the chance to use GenAI tools. It turned out that access to a GenAI assistant shortened the time they needed to complete their programming tasks by 56%. Brynjolfsson et al. [8] conducted an experiment among customer service workers in the Philippines. The opportunity to use GenAI tools resulted in the greater number of problems solved per hour.

3. The Impact of GenAI on the Labour Market – Study Conducted by NASK National Research Institute, Poland (NASK-PIB) and ILO

The method adopted in our study is an elaboration on the approach presented in 'Generative AI and Jobs: A global analysis of potential effects on job quantity and quality' [9]. The cited study assumes as its starting point the fact that every occupation consists of tasks assigned to it. Taking this into account, the researchers prepared a ChatGPT4 prompt (using the Application Programming Interface [API]) and asked the model to show the potential for automating a given task based on its linguistic description. ChatGPT assigned a value from 0 to 1 to each task, with 0 meaning automation is completely impossible, and 1 meaning that it is fully possible. The results were statistically elaborated with two measures defined: the average result (average of all tasks in an occupation) and standard deviation (distribution of automation results for tasks in an occupation).

As a result of this analysis, the authors suggested dividing the respective occupations into the following four groups, accounting for the probability of change due to AI tools becoming more and more popular:

- *Automation potential*: Occupations where most of today's tasks could theoretically be performed using GenAI – such occupations could potentially be automated without the need for human presence.
- *Augmentation potential*: Occupations where some of the tasks can be performed using GenAI, but most have to be performed by humans – such occupations may be augmented through GenAI, accelerating the performance of some tasks and providing more space for creative work for humans and new tasks.
- *Big unknown*: This category is between the *automation potential* and *augmentation potential*, representing jobs in which the balance of today's tasks hangs between those which can and those which cannot be performed with GenAI. As technologies develop and occupations evolve, this balance may shift, driving some occupations towards the *automation potential*, and some towards the *augmentation potential*.
- *Not affected*: Occupations in which most of the tasks cannot be performed using GenAI (e.g. physical tasks).

It is this method, modified, that is used precisely in the study performed by NASK-PIB) and ILO. Its most important assumption, like in Gmyrek et al. [9], is to use the assessment of the automation potential for tasks comprising a given occupation. The important difference lies in the fact that we divided the process of assigning an index value to the occupations into three stages: (1) assessing the potential for exposure of employee-performed tasks in a given occupation group (with 1600 respondents assessing the respective tasks participating at this stage of the study); (2) having the assessment verified by a group of experts; and (3) extrapolating the assessments onto 30,000 tasks representing 2500 occupations (using ChatGPT4).

The outcome is a comprehensive GenAI occupation exposure index that:

1. is based on a two-tier human assessment (workers in a given occupation and labour market experts);
2. is adapted to the national classification of occupations in Poland (the Polish Classification of Occupations and Specialisations [KZIS]), and, most of all, is based on the Polish linguistic descriptions of the tasks; and

3. covers all occupations on a six-digit International Standard Classification of Occupations (ISCO) level, that is a total of 2500 occupations (to be estimated by ChatGPT4).

Such an index enables a reliable estimation of the impact that GenAI has on the Polish labour market and calculate the population whose occupation is 'under threat' from the constant development of AI tools. This makes it possible to prepare macroeconomic forecasts to be used in estimating the general impact of GenAI on Polish economy. Additionally, the intention of the project's authors is to publish these data so that every economically active person may learn the forecast for their occupation and use this knowledge in further planning their careers.

References

- [1] R. Włoch, K. Śledziewska, S. Rozynek S. "Who's afraid of automation? Examining determinants of fear of automation in six European countries," *Technology in Society*, Vol. 81, p. 102782, 2025. doi: [10.1016/j.techsoc.2024.102782](https://doi.org/10.1016/j.techsoc.2024.102782).
- [2] P. Gmyrek, K. Kamiński, A. Ładna, K. Roślaniec, M. Troszyński, F. Konopczyński, J. Berg, B. Nafradi (2024). Raport Otwarcia. Potencjalny wpływ Generatywnej Sztucznej Inteligencji na ilość i jakość zatrudnienia w Polsce. [Online]. Available: <https://www.gov.pl/attachment/26a22f10-c8d1-4923-b747-0b02bc278a90> [Accessed: Dec. 30, 2024].
- [3] E. Felten, M. Raj, R. Seamans, "Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses," *Strategic Management Journal*, vol. 42, no. 12, pp. 2195–2217, 2021. doi: [10.1002/smj.3286](https://doi.org/10.1002/smj.3286).
- [4] E. Felten, M. Raj, R. Seamans. (2023). *How will language models like ChatGPT affect occupations and industries?* SSRN. Available: <https://ssrn.com/abstract=4375268> 2023. arXiv preprint arXiv:2303.01157. doi: [10.2139/ssrn.4375268](https://doi.org/10.2139/ssrn.4375268).
- [5] K. Korgul, L. Witczak, I. Święcicki. *AI na polskim rynku pracy*. Warsaw: Polski Instytut Ekonomiczny, 2024.
- [6] D. Acemoğlu, D. Autor, J. Hazell, P. Restrepo, "Artificial intelligence and jobs: Evidence from online vacancies," *Journal of Labor Economics*, vol. 40, no. S1, pp. 293–340, 2022.
- [7] S. Peng, E. Kalliamvakou, P. Cihon, M. Demirer. The impact of AI on developer productivity: Evidence from GitHub Copilot. 2023. arXiv preprint arXiv:2302.06590v1. doi: [10.48550/arXiv.2302.06590](https://doi.org/10.48550/arXiv.2302.06590).
- [8] E. Brynjolfsson, D. Li, L.R. Raymond, *Generative AI at work*, Working Paper Series, 31161. Cambridge, MA: National Bureau of Economic Research, 2023. doi: [10.3386/w31161](https://doi.org/10.3386/w31161).
- [9] P. Gmyrek, J. Berg, D. Bescond. *Generative AI and jobs: A global analysis of potential effects on job quantity and quality*, ILO working paper 96. Geneva: ILO, 2023.