NOVA IMS Information Management School

MGI

Master Degree Program in
Information Management

AI-DRIVEN CYBERSECURITY: A HUMAN-IN-THE-LOOP APPROACH

Caroline Helen da Silva

Master Thesis Dissertation

presented as partial requirement for obtaining the Master Degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

AI-DRIVEN CYBERSECURITY: A HUMAN-IN-THE-LOOP APPROACH

by

Caroline Helen da Silva

Master Thesis presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Information Systems Management.

Supervised by

Ph.D. Mijail Naranjo-Zolotov

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Caroline Helen da Silva

Lisbon, 2024

DEDICATION

To Naomi Utsunomiya, throughout this entire journey, your belief in me, unconditional support, and tête-à-tête encouragement moments were the tie that bound all of my efforts together.

ACKNOWLEDGEMENTS

I want to express my heartfelt gratitude to my NOVA IMS thesis supervisor, Ph.D. Mijail Naranjo-Zolotov. Your faith in my proposals, invaluable guidance, enduring patience, and constructive feedback throughout this research journey have been significantly treasured.

I also wish to extend my gratitude to my family for their unwavering and immeasurable support, Naomi, Roseli Sanae, Larissa, Michel, and Natan, thank you for being there for me during this journey, providing strength that propelled me until the completion of this work.

Finally, I want to acknowledge my profound appreciation for the enduring friendships made during the master's and also, all my professors at NOVA IMS. Their inspirational pieces of advice, dedication, and generosity in sharing knowledge, actively contributed to making this educational journey immensely rewarding.

ABSTRACT

Organizational cybersecurity is a complex socio-technical challenge that interlaces the protection of digital assets and cyberspace with human behavior vulnerabilities. The alarming rise in cyberattacks against the global critical infrastructure, combined with dreadful future projections, emphasizes the need for innovative cybersecurity strategies, such as to adopt an AI-driven cybersecurity approach, leveraging Artificial Intelligence (AI) techniques to fortify and proactively strength organizational cybersecurity solutions. However, AI advancements are also becoming increasingly accessible to malicious actors, intensifying cybersecurity challenges. To address that, the often-overlooked Human-in-the-loop (HITL) approach has the potential to significantly improve measures, rising above the limitations in automation and AI-driven cybersecurity techniques. This research path starts with a systematic literature review of cybersecurity, AI, HITL and the socio-technical approach, followed by an instrumental case study analysis of the destructive NotPetya cyberattack, that subsequently, led to the development of the AI-HITL dimensions and finally the proposition of the AI-HITL approach for cybersecurity, which allows customization of strategies to meet specific organizational needs, and provides the foundation for future research efforts. In the AI-HITL approach, complementarity, achieved through the distinct integration of AI and HITL capabilities, advances cybersecurity measures, facilitates resourcefulness and ensures a higher level of security and readiness to effectively combat cyber threats in this constantly evolving domain.

KEYWORDS:

Cybersecurity, Artificial intelligence, AI-Driven Cybersecurity, Human-in-the-loop, Social-technical systems, AI capabilities, HITL capabilities, AI-HITL approach

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

Statement of Integrity i
Dedication ii
Acknowledgements iii
Abstract iv
List of Figures vi
List of Tables vii
List of Abbreviations and Acronyms viii
1. Introduction 1
    1.1. Research Approach 2
2. Theoretical Background 3
    2.1. AI-Driven Cybersecurity 3
        2.1.1. Cybersecurity Teams 6
    2.2. The Socio-Technical and Human-In-The-Loop Approach for Cybersecurity 6
    2.3. Solutions to Cybersecurity Challenges 8
        2.3.1. Hybrid Intelligence 8
        2.3.2. Resilience 10
        2.3.3. Governance 12
    2.4. Conceptual Framework: Integrating HITL in AI-Driven Cybersecurity 14
3. Methodologies 16
4. Case Study Analysis 20
    4.1. The Notpetya Cyberattack 20
5. Results and Discussion 24
    5.1. The AI-HITL-Driven Cybersecurity Approach 24
    5.2. The AI-HITL-Driven Cybersecurity Core Components 25
    5.3. The Dimensions of the AI-HITL Approach in Cybersecurity 26
        5.3.1. AI-HITL Intelligence 26
        5.3.2. AI-HITL Resilience 26
        5.3.3. AI-HITL Governance 27
    5.4. The AI-HITL in Cybersecurity Teams 28
    5.5. The AI-HITL Approach to the Notpetya 30
6. Conclusions and Future Works 33
Bibliographical References 34

LIST OF FIGURES

Figure 1 - Cybersecurity STS dimension attributes 7
Figure 2 - Complementary strengths of humans and machines 9
Figure 3 - The resilience paradigm 11
Figure 4 - The emerging triad 14
Figure 5 - Dimensions of AI-HITL in cybersecurity capabilities 15
Figure 6 - SLR PRISMA Process 17
Figure 7 - Socio-technical IS design theory development 18
Figure 8 - Instrumental Case Study Process 19
Figure 9 - The AI-HITL in Cybersecurity 24
Figure 10 - AI-HITL Intelligence 26
Figure 11 - AI-HITL Resilience 27
Figure 12 - AI-HITL Governance 28
Figure 13 - The AI-HITL approach to cybersecurity teams 29

LIST OF TABLES

Table 1 - AI techniques for Cybersecurity 5
Table 2 - The Search Query..... 16
Table 3 - SLR Inclusion and Quality criteria 17
Table 4 - Vulnerable points from the NotPetya Case Study..... 22
Table 5 - The AI-HITL Core Components..... 25
Table 6 - The AI-HITL approach to the NotPetya..... 31

LIST OF ABBREVIATIONS AND ACRONYMS

ACD	Automated Cyber Defense
AI	Artificial Intelligence
AIC	Artificial Intelligence Capabilities
AI-HITL	Artificial Intelligence and Human-in-the-loop
AI RMF	NIST AI Risk Management Framework
APTs	Advanced Persistent Threats
ASI	Artificial Swarm Intelligence
CSF	NIST Cybersecurity Framework
DL	Deep Learning
ENISA	European Union Agency for Cybersecurity
GenAI	Generative Artificial Intelligence
HITL	Human-in-the-loop
HITLC	Human-in-the-loop Capabilities
ML	Machine Learning
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
SLR	Systematic Literature Review
STISD	Socio-Technical IS Design Science Research
STS	Socio-Technical Systems
XAI	Explainable AI

1. INTRODUCTION

Cybersecurity presents a multifaceted socio-technical dilemma (de Bruijn & Janssen, 2017) as it involves processes, methods and human behavior in safeguarding the cyberspace and digital assets from unauthorized access and cyberattacks (Shaukat et al., 2020; Zeadally et al., 2020). With an increase of 30% from 2022, Forescout Research – Vedere Labs recorded, between January and December of 2023, more than 420 million cyberattacks to the world's critical infrastructure, with an average of 13 attacks happening every second (Forescout Vedere Labs, 2024) and the future predictions are not promising, with an estimated cybercrime cost of $10.5 trillion by 2025 (Astra, 2024).

To ensure a smooth business operation and reliable delivery of the designed services, companies current have at their disposal diverse cybersecurity techniques, tools and a range of security measures to fortify and shield systems from several cyber dangers (Shaukat et al., 2020; Zhang et al., 2022). Gaining significance over the years, the adoption of Artificial Intelligence (AI) for addressing cyber threats has been offering significant benefits to businesses across the board (Kaur et al., 2023; Wiafe et al., 2020). Nevertheless, while amplifying capabilities in defending cyberspace, AI also can serve as a potential tool for malicious intent, empowering cybercriminals to bypass the protective mechanisms in place (Shaukat et al., 2020; Zeadally et al., 2020).

Beyond the capability of any current AI and automated technologies, human involvement, also known as the Human-in-the-loop (HITL) approach, is irreplaceable for increasing cybersecurity countermeasures, as humans still hold distinctive capabilities and contextual expertise that are critically indispensable for decision-making and proper governance of AI (Dellermann et al., 2019; Zhang et al., 2022; Zimmermann & Renaud, 2019). Supporting that, the European Union Agency for Cybersecurity (ENISA) released a study titled *"Artificial Intelligence and Cybersecurity Research"* on its official website. In this study, ENISA identified research gaps related to AI that hold significant relevance for cybersecurity applications, one of which is labeled "*Bringing 'humans into the loop'*" (ENISA, 2023, p. 33). Additionally, concerning the need for further research, ENISA highlighted in the area of AI training that "*the best AI requires data scientists, statistics, and as much human input as possible*" (ENISA, 2023, p. 36).

Likewise, The National Institute of Standards and Technology (NIST) released in February 26, 2024 an update from its renowned NIST Framework for Improving Critical Infrastructure Cybersecurity version 1.1 to the NIST Cybersecurity Framework (CSF) 2.0, introducing a new function, Govern, where HITL principles are actively integrated to guarantee a dynamic and responsive cybersecurity posture, facilitate effective communication and enhance decision-making between executives, managers, and practitioners (National Institute of Standards and Technology, 2024). Furthermore, in 2023, NIST released The NIST AI Risk Management Framework (AI RMF), centered around four core functions: Govern, Map, Measure and Manage. The AI RMF recognizes the socio-technical essence of AI systems, acknowledging that

"they are influenced by societal dynamics and human behavior" (National Institute of Standards and Technology, 2023, p. 1).

However, as AI increasingly evolves as a critical component of cybersecurity, existing AI-driven cybersecurity research predominantly focus on enhancing AI technology, mainly through advancements in AI methodologies and algorithms, and often disregarding the critical importance of integrating humans-in-the-loop (Heyder et al., 2023; Kaur et al., 2023; Zhang et al., 2022).

1.1. RESEARCH APPROACH

To bridge this research gap, this study contributes to academic research by adopting a systematic literature review (SLR) approach (Kaur et al., 2023; Zhang et al., 2022), using as basis the PRISMA framework (Page et al., 2021) and the Socio-technical theory approach (Bostrom & Heinen, 1977b, 1977a; Carlsson et al., 2011; Taddeo, 2019), combined with an instrumental case study research (Cousin, 2005; Ebneyamini & Sadeghi Moghadam, 2018), to explore AI in cybersecurity, emphasizing the integration of HITL within organizational contexts, with the aim of addressing the following research questions:

- RQ1: How can we effectively integrate HITL for AI-driven cybersecurity?
- RQ2: How does the integration of HITL for AI-driven cybersecurity contribute to advancing cybersecurity measures in businesses?

Discussing AI-driven cybersecurity in this research, in the next chapter, we focus on three AI techniques: Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP), complemented by the socio-technical systems (STS) theory and the HITL approach. As of the conclusion of this research, available information does not indicate the existence of a thorough review addressing the current state-of-the-art research, which is integrating HITL into AI-driven cybersecurity, specifically aimed at augmenting organizations capabilities.

2. THEORETICAL BACKGROUND

2.1. AI-DRIVEN CYBERSECURITY

Cybersecurity threats, ever-changing complex hacking attacks, security issues and data breaches in today's digital age are becoming the norm rather than the exception, to the point where it has finally captured mainstream awareness and there are no indication of it slowing down anytime soon (Chowdhury & Gkioulos, 2021; de Bruijn & Janssen, 2017; Zimmermann & Renaud, 2019). In the past, the question centered around the possibility of an organization being compromised. Today, it has evolved into a question of when and to what extent (Jalali et al., 2019).

With AI being a large, multidisciplinary research area, it encompasses a range of technologies and applications (Kaur et al., 2023) that can be leveraged as a strategic competitive tool within organizational cybersecurity (Michael et al., 2023). As organizations are faced with rising complexities in cybercrimes, where each day the gravity of threats increases, with more sophisticated and unusual critical cyberattacks, AI is increasingly evolving as a critical component of cybersecurity for organizations of all sizes and sectors (Kaur et al., 2023; Wiafe et al., 2020).

In this degree, AI-driven cybersecurity is already a global pressing concern, urging new efforts to improve security measures in detecting, responding and even counterattacking to threats, as traditional security systems have innumerous limitations, such as being outdated and inefficient in detect multilayered and polymorphic attacks (Rana et al., 2022; Shaukat et al., 2020). Despite traditional machine learning (ML) techniques initially not being designed to be used in security, nowadays ML techniques have been widely applied within the field of cybersecurity in a wide range of cybersecurity systems to address critical threats such as phishing, malware and ransomware, that are ranking as the most significant forms of cyberattacks, and being recognized as major security threats for organizations (Almashhadani et al., 2022; Bertia et al., 2022; Shaukat et al., 2020; Zeadally et al., 2020).

On the defensive strategy, ML techniques can be employed to develop intelligent agents capable of distinguishing between illegitimate attack traffic and legitimate network activity. Then, sequence pattern mining techniques can be utilized to extract patterns from network data and feed it into ML classifiers to effectively detect and prevent ransomware attacks (Almashhadani et al., 2022; Dargahi et al., 2019; Zeadally et al., 2020). Also, ML has the capability to make predictions and decisions, learning from experiences and responding accordingly, with fast and more precise countermeasures that allow early identification and anticipation of cyberattacks (Shaukat et al., 2020; Zeadally et al., 2020).

Likewise, Deep learning (DL), that can be defined as a subset of ML, share the same ML techniques and tasks but with different capabilities (Neupane et al., 2022; Saha et al., 2021; Shaukat et al., 2020). In cybersecurity, DL techniques has been allowing machines to find

hidden correlations within input data, consequently producing more precise outcomes for risk planning and prediction (Zeadally et al., 2020). DL also can be an ally in a new concern for organizations, which is Advanced Persistent Threats (APTs), that are different from conventional network attacks, and it is known for their devastating capacity of damage. Employing complex attack methods to exploit vulnerabilities in network systems and protocols, and showing relatively few identifiable characteristics, these attacks allow access to sensitive data and disruptions in network operations (Al-Kadhimi et al., 2023; Al-Saraireh & Masarweh, 2022; Sharma et al., 2023). Multiple ML classifiers can be used to identify APTs based on established patterns. DL improves this, with its capability to rapidly and precisely process extensive volumes of data, extracting complex features from unprocessed data without necessitating feature manipulation, it accelerates the detection of APTs (Al-Kadhimi et al., 2023; Al-Saraireh & Masarweh, 2022; Jia et al., 2023; Shaukat et al., 2020).

Also, on APT attacks, sophisticated zero-day malware is a core element (Sharma et al., 2023). In cybersecurity, the term zero-day is used to define new and unknown, not disclosed before (Ahmad et al., 2023). For example, even the latest security software may not spot code concealed using evolved techniques for code obfuscation (Zeadally et al., 2020). To help in the identification of zero-day attacks, DL can autonomously identify irregularities with limited supervision, by examining complex data and learn from past attack patterns (Ahmad et al., 2023; Keshk et al., 2023). Yet, despite DL being a rapidly evolving theory, it still remains complex and not transparent enough (Shaukat et al., 2020).

Moreover, Natural language processing (NLP) can provide a proactive threat intelligence, where NLP can help in analyzing email content or communication related to code delivery, recognizing subtle linguistic cues or anomalies that indicate potential malicious intentions (Kaur et al., 2023; Shaukat et al., 2020). With NLP, it is also possible to predict cyber threats using sentiment analysis, forecasting cyber threats several weeks in advance of their occurrence (Deb et al., 2018).

To further showcase the multitude of potential AI techniques in the cybersecurity domain, Table 1 has been adapted from the taxonomy proposed by Kaur et al. (2023). These techniques emerged from cybersecurity functions, ranging from the utilization of AI capabilities to prevent security breaches, to proactively detecting and counterattacking threats (Kaur et al., 2023). Adjustments were made to best fit this research on the correlations of cybersecurity activities and AI-driven applicable techniques strategies.

Therefore, it is also important to take in consideration that, while researchers and experts actively explore and implement AI techniques to strengthen cybersecurity countermeasure capabilities, such as enhancing system resilience through user authentication, device verification, asset validation, and monitoring user activities, these efforts also introduce nuanced vulnerabilities. For instance, cybercriminals exploiting AI to sophisticate their hiding strategies and power their attacks, by poisoning data and avoiding detection (Gupta et al., 2023; Jalali et al., 2019; Kaur et al., 2023; Neupane et al., 2022; Zeadally et al., 2020).

Table 1 - AI techniques for Cybersecurity – Adapted from (Kaur et al., 2023)

Cybersecurity Activities	AI Techniques Strategies
Risk Assessment	▪ *Automated IT Asset Configuration, Control, and Inventory Risk Assessment* ▪ *Automated Risk Analysis, Business Impact Analysis, and Assessment* ▪ *Automated Risk Policy Enforcement and Risk Documentation Analysis* ▪ *Automated Supply Chain Security Risk Assessment* ▪ *Automated Threat Hunting, Vulnerability Identification, and Assessment* ▪ *Decision Support for Risk Planning and Predictive Intelligence* ▪ *Dynamic Risk Scoring*
Safeguard	▪ *Adaptive Security Monitoring and Control: Anomaly Detection and Behavior Analytics* ▪ *AI-Enhanced Vulnerability Management Plan* ▪ *AI-Powered Backup* ▪ *AI-Powered Honeypots* ▪ *AI-Supported user and device authentication* ▪ *Anti-virus/Anti-malware* ▪ *Automated Access control* ▪ *Automated Assessment of Threat Intelligence sources* ▪ *Data leakage prevention* ▪ *Dark Web investigation* ▪ *Intelligent E-mail protection* ▪ *Intrusion Detection System and Intrusion Prevention System* ▪ *Log Analysis* ▪ *Malicious domain blocking and Reporting* ▪ *Protection by deception*
Mitigation Response	▪ *AI-Driven Business Continuity Planning* ▪ *AI-Guided Incident Response Playbooks* ▪ *Alert Triage and Dynamic Case Management* ▪ *Automated Incident Characterization* ▪ *Automated Isolation* ▪ *Automated Remediation* ▪ *Automated Responsibility Allocation* ▪ *Collaboration Support System* ▪ *Forensic Analysis* ▪ *Long-term Improvements*
Recovery	▪ *AI-Enabled Data Reconstruction* ▪ *AI-Enhanced Incident Review and Lessons Learned* ▪ *Analysis and Aggregation of Incident Reports* ▪ *Automated Backup and Restoration* ▪ *Information Sharing Property Platform*

2.1.1. Cybersecurity Teams

Following Wiafe et al. (2020) suggestions: *"Cyber defense mechanisms must be i) increasingly intelligent, ii) more flexible, and iii) robust enough to detect a variety of threats and mitigate against them"* (p. 1). To achieve this in organizations, resilience is a key word, and the best way to test it is through cybersecurity teams real-life simulations, also known as competitions (Chindrus & Caruntu, 2023).

For example, offensive security professionals, also called Red teams, play a role similar to hostile forces, where they conduct complex attacks to take advantage of weak spots. On the other side of response, defensive security professionals, also called Blue teams, identify, counter, and defend against the attacks. Additionally, there is also Purple teams, which integrates both Blue and Red teams, and provide a comprehensive evaluation of technological security (Chindrus & Caruntu, 2023; Halisdemir et al., 2022; Titus & Russell, 2023).

Particularly with an AI-driven cybersecurity approach, Vyas et al. (2023) study proposes "*Automated Cyber Defence"* (ACD), which emphasizes on automated decision-making tools to enhance cyber defense strategies and deal with cyber threats. The ACD concept is in alignment with statements from both corporate and state entities, predicting that AI will soon take a leading role in cybersecurity. Regarding cybersecurity teams, similar to the human teams but powered by AI algorithms, the purpose of automated blue teaming is to continuously safeguard against harmful processes infiltrating the network system through its endpoints. Complementary automated red teaming aims to ensure the resilience of the automated blue teaming by identifying vulnerabilities and exploiting opportunities in the network system (Vyas et al., 2023).

Nevertheless, it is important to consider that even though AI technologies in cybersecurity teams offer significant help in identifying vulnerabilities and implementing innovative safeguards, HITL approaches are also essential to harmonize technical, ethical, and societal considerations in human-AI teaming, where governance is a part of the strategy development process, not an afterthought. By integrating socio-technical considerations into the design process, more robust, safe, and accountable cybersecurity solutions can be created (Hauptman et al., 2023; Malatji et al., 2020; Michael et al., 2023; Titus & Russell, 2023).

2.2. The Socio-Technical and Human-in-the-loop Approach for Cybersecurity

To fully employ the potential of AI-driven for cybersecurity it is necessary to understand the paradoxical nature it embodies, recognizing the intricacies and complexities present currently in socio-technical systems (STS) (Michael et al., 2023). Socio-technical systems can be defined as *"essentially what cyber systems are, are made up of multiple interconnected components, including human actors, technology and processes*" (Zimmermann & Renaud, 2019, p. 3).

Malatji et al. (2020) study summarized real-life examples of destructive data breaches and cyberattacks, facilitated by human mistakes or actions. They combined different authors perspectives and suggested the *"Cybersecurity STS dimension attributes".* Their conclusion is that regardless humans still being the biggest security risk in organizations, effective protection of assets requires a smooth integration of governance and compliance security frameworks that consider both technical and non-technical security controls (Malatji et al., 2020), as it is illustrated on Figure 1.

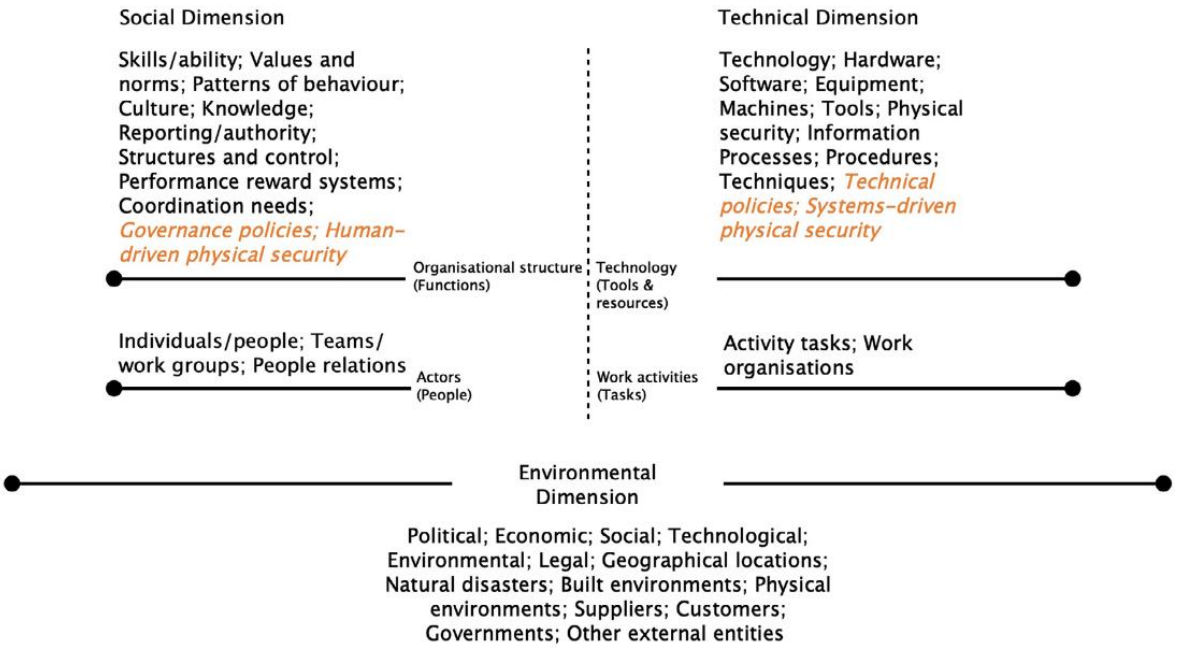

Figure 1 - Cybersecurity STS dimension attributes (Malatji et al., 2020)

Furthermore, established cybersecurity frameworks that integrates socio-technical perspectives still remains an evolving area, and an overreliance on AI in cybersecurity may increase the risks of complex socio-technical systems, as concentrating solely on AI limits organizations to a comprehensive understanding of important social components, such as governance practices and structure in general, that directly influences cybersecurity countermeasures capabilities, as the outcomes of organizational processes results from the collaborative interaction between these components (Malatji et al., 2020; Michael et al., 2023; Taddeo et al., 2023; Zimmermann & Renaud, 2019).

While still there is not a single solution, a socio-technical strategy that englobes this multiple interactions and complexity, rather than fixating exclusively on the attractive technological solutions, can be a differential for organizations customized strategies in cybersecurity, as it can help manage technical systems in a more effective way while addressing the multifaceted nature of organizational dynamics (de Bruijn & Janssen, 2017; Malatji et al., 2020; Taddeo et al., 2023; Zimmermann & Renaud, 2019). Likewise, ongoing discussions on AI versus humans or limiting HITL only hold the potential for achieving solutions that blend their combined

capabilities, as considering both technological and human-centric factors can improve the fostering of a security-oriented organizational culture aligned with the established organizational values (Cohen et al., 2023; Taddeo et al., 2023).

Even more, this multitude of possibilities in AI and cybersecurity presents an opportunity for Information Systems (IS) field to start acknowledging the benefits of socio-technical thinking for organizations (Berente et al., 2021; Malatji et al., 2020), specifically in cybersecurity, with its intricate complexity, interconnected nature, and ever-evolving challenges (Taddeo et al., 2023; Zimmermann & Renaud, 2019). For example, for organizations, ML techniques can help alleviate the shortage of qualified professionals that have expertise in specialized technologies for detecting cyber threats (Shaukat et al., 2020). Yet, in situations that requires human knowledge of the context or the exclusively human ability of assigning meaning to information, it is still imperative the need for humans to remain in the decision-making looping (Michael et al., 2023; Zanzotto, 2019).

With the adoption of the HITL approach it is possible to acknowledge the strategic value of incorporating human input into AI systems, for example, in decision-making processes is vital for user acceptance and organizational reflexivity, enabling organizations to reflect, adapt, learn and evolve from experiences (Baroni et al., 2022; Grønsund & Aanestad, 2020; Metcalf et al., 2019). It is already a reality that AI-based algorithms have high precision and can operate autonomously at brief intervals, but also the HITL approach is needed in organizations when human intervention might be fundamental to avoid cyber disasters, such as, when identifying and responding to scenarios that machines may overlook or when handling exception-related alerts beyond their operational capacity (Michael et al., 2023; Zimmermann & Renaud, 2019).

However, it is also important to acknowledge that humans still are portraited as the "Achilles heel of cybersecurity". Yet, at the same time that humans are not infallible, AI is not bulletproof. An excessive reliance on AI in cybersecurity can give a false sense of protection, increasing vulnerability. On the opposite, an insufficient use of AI techniques, that could increase the fighting capabilities against cyber threats, can also bring severe consequences in an organizational context (Michael et al., 2023; Sharma et al., 2023; Taddeo et al., 2023).

2.3. Solutions to Cybersecurity Challenges

2.3.1. Hybrid Intelligence

Believed to be a solution to overcome the hurdles mentioned before, the integration of human and artificial intelligence, what is also called Hybrid Intelligence, complement humans and machines strengths and allow the achievement of advanced levels of performance (Figure 2) (Dellermann et al., 2019). Human intelligence can be defined as the mental capabilities inherent in human beings. AI, as it name says, is artificial intelligence, capable of emulating human thinking, processing, and learning from external data to transform it into knowledge. And the term hybrid intelligence is linked to fostering continuous improvement through the

exchange of intelligence, knowledge, and skills between humans and AI (Dellermann et al., 2019; Johnson et al., 2022; Taeihagh, 2021).

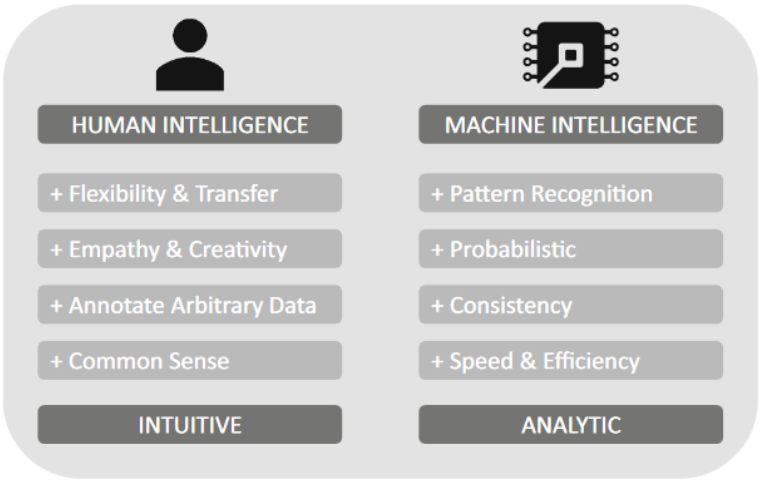

Figure 2 - Complementary strengths of humans and machines (Dellermann et al., 2019)

In Huang & Rust (2018) study, highlighting human intelligence and AI, they proposed a five-stage framework for the potential replacement or complete integration of humans by AI in organizations of different sectors. They suggested the four intelligences in AI development, arranged based on the level of complexity for AI to master: mechanical, analytical, intuitive, and empathetic. Mechanical intelligence involves AI automated execution of routine tasks with limited adaptability. Analytical intelligence centers around AI processing information to solve problems and gaining knowledge from the process, such as ML. Intuitive intelligence focus on AI that mimics human cognitive abilities for creative thinking and adaptability to new scenarios. And empathetic intelligence, encompassing AI recognizing and understanding emotions. Their forecast is that AI will master all these four intelligences and achieve a level of intelligence comparable to humans (Huang & Rust, 2018).

Meanwhile, approaching the intelligence of human groups, Metcalf et al. (2019) explore the hybrid intelligence between groups of humans and machines, presented as a key to improve group decision-making. The authors suggested exploring the utilization of the HITL approach in ML, particularly focusing on Artificial Swarm Intelligence (ASI). The authors argue for the unique advantages of ASI, emphasizing its ability to access the unique human tacit knowledge, one significant limitation of ML, and act as a collaborative brain, what they call "*kind of Supermind*". The paper findings suggest that ASI can be leveraged in decision-making, where all swarm expertise is reachable to managers or organizations. Their conclusion is that ASI amplifies the intelligence of a group and surpasses the intelligence of ML (Metcalf et al., 2019).

As another example, Chandra et al. (2022) study presented a theoretical framework that identified three AI competencies: cognitive, relational, and emotional, focused on conversational AI that can enhance user engagement. While our focus is not on conversational AI, these competencies can be linked to the hybrid intelligence approach for organizational

cybersecurity. As definitions, cognitive competency processes information for event interpretation and task optimization, while relational competency encourages cooperation and fosters communication, and emotional competency involves displaying emotional states and empathy. By applying these competencies into the hybrid intelligence approach, cybersecurity defense techniques for organizations can become more adaptive, resilient, and responsive, for example, cognitive competency can enhance threat analysis, identifying sophisticated cyber threats and vulnerabilities. Relational competency can contribute to effective team collaboration, facilitating the threat response process. And emotional competency can be a differential in understanding the emotional context of cyber threats to prevent them (Chandra et al., 2022).

With cybersecurity solutions moving from AI agents that act as humans to AI agents that really think as human, the human intelligence can be leveraged in combining the available information from various machines and using its unique abilities of comprehension to better respond to ever-changing challenges (Michael et al., 2023; Nunes et al., 2015; Zeadally et al., 2020; Zhang et al., 2022). Currently, AI developers prioritize when simulating the human mind thinking process, the immediate cognitive response, but it is also necessary to consider that the human brain has a dual cognitive thinking system, a rapid automated thinking system and a slower and more reflective thinking system (Chandra et al., 2022; Huang & Rust, 2018). Also, the decision rules AI learns from data are different from human decision rules, as humans make choices with probabilities, using their capacity of self-reflection and social reasoning to provide insights. While AI always goes for the most likely option, with advanced AI systems learning from experiences rather than just following strict rules set by humans (Fügener et al., 2021; Namvar et al., 2023).

Ultimately, it is not only important to integrate human intelligence but also human knowledge and competencies (Chandra et al., 2022; Huang & Rust, 2018; Johnson et al., 2022). Specifically addressing decision-making, it differs between humans and AI, as for humans it combines key aspects that contribute to better outcomes, such as capacity of innovate, meaningful ideas, mathematical proficiency, thoughtful reasoning, and memory skills. In contrast, AI achieves predictions through meticulous methods, such as rule-based deductions or statistical algorithms, ensuring decision-making within acceptable limits of bias and risk (Caldwell et al., 2022; Dellermann et al., 2019; Nunes et al., 2015).

2.3.2. Resilience

It is already a reality that AI techniques and human–machine hybrid methods can combine the quality of human creativity together with the high performance of the algorithm to achieve high-quality results (Wu et al., 2022). Likewise, as AI solutions and techniques evolve, resilience is also critical (Shaukat et al., 2020). Talking about different resilience perspectives, contextual understanding, humans perceive resilience in an agile approach, focusing on response and recover (Blair et al., 2019; Zimmermann & Renaud, 2019), while AI resilience focuses on the robustness and performance of AI systems themselves (Nunes et al., 2015).

In the context of cybersecurity, another used term is cyber resilience, which in a human-oriented perspective usually denotes the ability to fast recover to pre-event levels of functioning after a disturbance. While in a machine-oriented view is the insurance of systems capacity to maintain essential functions despite threats, conditions being changed and possible incidents. In that way, AI collaboration in enhancing cyber resilience in organizations is through its solutions (Domínguez-Dorado et al., 2022; Kaur et al., 2023; Zimmermann & Renaud, 2019). An example of a solution is resilient DL models that can endure noise and adversarial instances in cybersecurity (Shaukat et al., 2020). However, DL techniques, for example, can have a considerable rate of false positives (Keshk et al., 2023) and it is where HITL is needed, with experts filtering the unlabeled set, calibrating and re-training the target model to improve results (Wu et al., 2022; Zhang et al., 2022).

Using as example Wu et al. (2022), their findings through the investigation of HITL for ML, gathered that a system incorporating HITL outperforms those without it. Additionally, the authors proposed NLP approaches that employ human intelligence for training and drawing inferences from experimental results in a continuous executive loop, where this collaboration enhances accuracy and robustness in the NLP system. Within the suggested collaborations between AI algorithms and humans, human knowledge was actively incorporated in the proposals and successfully improved models strengthening and resilience (Wu et al., 2022).

Similarly, Nunes et al. (2015) explored cyber-physical systems that incorporated HITL and proposed the resilience paradigm (Figure 3), defining resilience as: "*a combination of a number of features, which include dependability, security and privacy, as well as the required means to provide overall robustness and performance to the system*" (p. 14). However, beyond the plenty of differences in decision-making, resilience, processing of information and knowledge between machines and humans, to understand machine decisions in order to guarantee trustworthiness is very challenging to humans (Caldwell et al., 2022; Coombs et al., 2021; Nunes et al., 2015; Shaukat et al., 2020).

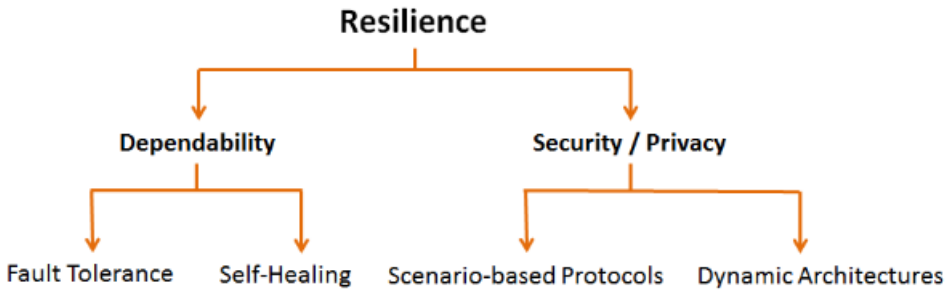

Figure 3 - The resilience paradigm (Nunes et al., 2015)

Due to the intricate and adaptive nature of ML, DL and NLP it is still often difficult for humans to figure out the rationality behind the algorithms decision-making (Taeihagh, 2021). Specifically in DL, not being able to break down and track-back the rationality into easy to

understand parts, what is called decomposability, is not only a challenge, but also a risk for organizations (Coombs et al., 2021; Neupane et al., 2022). Discussing NLP, trustworthiness is an essential impact factor, particularly in text classification, and must be integral where text meaning are correlated directly with detection results (Shaukat et al., 2020). Nevertheless, it is usually critical for organizations that algorithms remain opaque to not give breach opportunities for cyberattacks and expose sensitive data (Taeihagh, 2021).

2.3.3. Governance

To address the complexity of social-technical systems, a solution is necessary to balance humans, machines, and corporate governance policy (Kane et al., 2021; Minkkinen & Mäntymäki, 2023; Wiafe et al., 2020). For cybersecurity, according to Kaur et al. (2023) governance helps in assessing the organizations responsibilities and providing cyber risks information to the management. Their definition is that "*governance involves the policies, procedures and processes for understanding the environmental and operational requirements, and monitoring the regulatory requirements of the organization"* (Kaur et al., 2023, p. 8).

Furthermore, specific and tailored governance guidelines need to be created through a proactive and unified approach throughout the organization, focusing on addressing not only cybersecurity risks, but also business continuity, incident response and recovery. AI can help in implementing and enforcing policies, as well as, incorporating automated processes for the extraction of critical cyber risk indicators (Domínguez-Dorado et al., 2022; Kaur et al., 2023). For example, transforming critical risk indicators, such as, the presence of unpatched systems or attempted breaches, into actionable knowledge through automatic extraction is essential when dealing and mitigating cybersecurity threats (Cram et al., 2017; Kaur et al., 2023). Yet, current governance and regulatory cybersecurity frameworks are inadequate in addressing socio-technical challenges posed by AI and employees risky behaviors (Cram et al., 2017; Taeihagh, 2021).

AI systems, by their nature, are a target for ill-intentioned actors, and for benevolent and authorized actors that can commit an error or mistake. If this socio-technical complexity is oversighted, in this reality of daily use of AI powered tools in organizations, it may lead to the potential misuse of AI in cybersecurity (Gupta et al., 2023; Kane et al., 2021; Malatji et al., 2020; Taddeo et al., 2023). In Gupta et al. (2023) study, they discussed about the social, legal, and ethical implications of the use of Generative AI (GenAI) and its impact on cybersecurity and privacy, where they researched about how different GenAI models, like ChatGPT, can be used for malicious purposes in cyber offense, for example, exploiting hallucinations, automated hacking, empowering social engineering, ransomware and malware code generation. And also, as it can be an ally for cyber defense, to enhance threat intelligence, defense automation on detection, reporting, incident response, and other existing utilities. The governance of these different aspects can be a direction to ensure safety and compliance in this new reality (Gupta et al., 2023).

Even more, regarding our focus on ML, DL and NLP, governance is necessary to deal with the unpredictability and complexity of AI advancements. To enhance inclusivity and diversity in AI governance is important that organizations shape their AI policies prospect considering all stakeholders respective participation. Innovative governance methodologies may be the way to face the innumerous troubles when formulating and executing effective policies in the complex AI for cybersecurity landscape (Kane et al., 2021; Taddeo et al., 2023; Taeihagh, 2021).

Therefore, strategic management has the power to shape human-AI interaction, covering opacity, unfairness, and preventive, reactive and ongoing processes (Heyder et al., 2023). By creating specific and targeted AI governance that aligns with cybersecurity corporate governance, it is possible to understand and deal with the risks and uncertainties of AI systems applications in cybersecurity (Heyder et al., 2023; Kaur et al., 2023; Taeihagh, 2021).

However, only focusing on governance and compliance does not seem to be sufficient anymore in this ever-evolving reality, as cybercriminals constantly change their techniques and use new technology at an unprecedented speed. Also, counting only on policies to prevent errors and mistakes and regulate human behavior fails to recognize this new reality (Malatji et al., 2020; Zimmermann & Renaud, 2019). Evaluating the efficiency of both technical and non-technical security measures ensures optimized protection of assets, enhancing guidelines that align with specific organizational requirements and regulatory compliance (Kaur et al., 2023; Malatji et al., 2020; Minkkinen & Mäntymäki, 2023).

Besides, to effectively supervise, assess, update, and authorize the compilation of cybersecurity policies, a more intricate governance system is necessary. In the same way, relating to AI, to be able to effectively supervise it, is necessary to comprehend the logic behind AI provided outputs, and to address that, there is explainable AI (XAI), explaining how the thinking behind the AI decision-making was. Although there are complex and layered algorithms, for example DL, that are harder to follow through the decision-making path, it is essential to ensure that AI systems uncertainty is measured and available to improve AI governance. At the same time, AI systems must provide traceability of the final results and support the integration of different forms of human participation. Not only algorithms are complex, but so are humans, and its behaviors and interactions affect explainability (Caldwell et al., 2022; Coombs et al., 2021; Heyder et al., 2023; Taddeo et al., 2023).

Most of the AI-related cybersecurity policies only aim at governing AI in security and defense domains, ignoring the socio-technical interconnected relationship with AI systems, but no matter how robust and complete a system can be, even if it is safe from external threats, it can be vulnerable to insider threats. If strategic management is positioned equally important as algorithm performance, it offers an augmented cybersecurity protection and improve results in corporate governance strategies (Caldwell et al., 2022; Grønsund & Aanestad, 2020; Heyder et al., 2023; Malatji et al., 2020; Michael et al., 2023; Taddeo, 2019; Taddeo et al., 2023).

2.4. Conceptual Framework: Integrating HITL in AI-Driven Cybersecurity

Building upon the findings from the SLR, specifically the solutions to cybersecurity challenges, which extensively addressed cybersecurity, AI and HITL integration specifics using a socio-technical approach, led to the development of a conceptual framework (Al-Kadhimi et al., 2023). To start, this research proposes the emerging triad (Figure 4), being AI integrated with HITL (AI-HITL) for cybersecurity.

The term triad is used to emphasize the significance and relationship of the promising HITL potential when combined with AI techniques capabilities for cybersecurity application, where the perfect balance between effective performance, accuracy, and trustworthiness can be found in the collaboration between human expertise and AI techniques in systems that leverage HITL (Cohen et al., 2023; Michael et al., 2023).

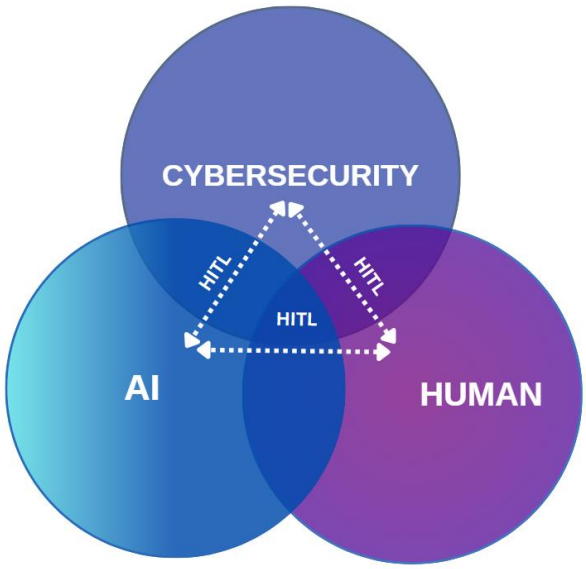

Figure 4 - The emerging triad (own construction)

Besides the HITL correlations between AI, cybersecurity, and human, addressing inside each component of the emerging triad we propose the dimensions of AI-HITL in cybersecurity countermeasures capabilities (Figure 5), encompassing three dimensions: Intelligence, Resilience and Governance.

These dimensions are recapitulated as design propositions aligned with our thick descriptions (Carlsson et al., 2011), where the combined potential, guided by the HITL approach (Michael et al., 2023) across these three key dimensions, enables a more intelligent and effective response to the always changing landscape of cybersecurity threats, foments and improves cybersecurity resilience, and ensures the implementation of advanced and beneficial governance for cybersecurity.

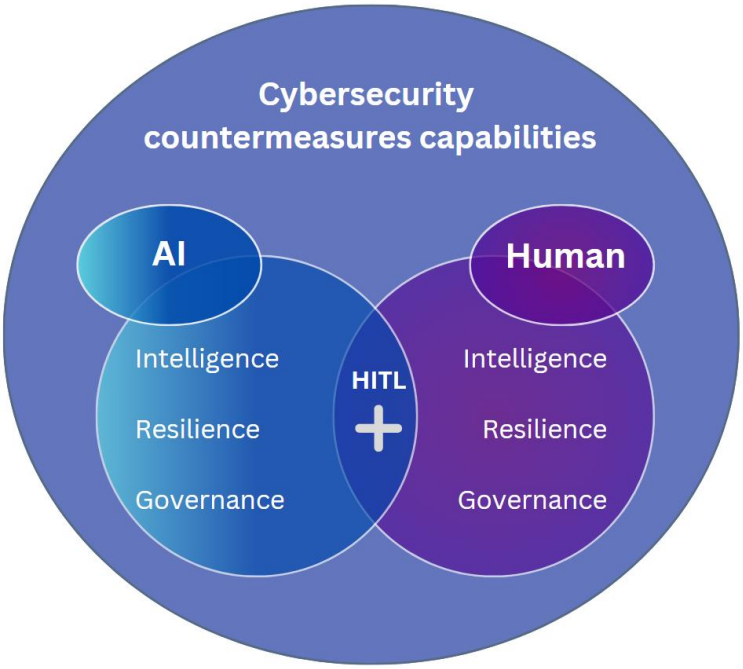

Figure 5 - Dimensions of AI-HITL in cybersecurity capabilities (own construction)

3. METHODOLOGIES

In this research methodology, the Systematic Literature Review (SLR) aims to explore, analyze and interpret all the relevant available research to spotlight research gaps and future trends (Kaur et al., 2023; Page et al., 2021). Especially dedicated to exploring AI-driven cybersecurity and emphasizing the integration of HITL within organizational contexts.

Recognizing the crucial role of the search strategy in any research (Al-Kadhimi et al., 2023), this SLR strategy is organized based on the PRISMA approach (Page et al., 2021) to align with our research questions. Therefore, to identify the most relevant and recent existing studies, different platforms were used, using a set of comprehensive keywords and Boolean operators, as presented in Table 2.

Table 2 - The Search Query (own construction)

Databases	Keywords and Boolean operators
ScienceDirect, Scopus, IEEE Xplore, EBSCOhost, and Google Scholar.	– (“artificial intelligence” OR “AI” OR “machine learning”) AND (“cybersecurity” OR “cyber security”) AND (“human-in-the-loop” OR “human in the loop” OR “HITL”), – (“artificial intelligence” OR “AI” OR “machine learning”) AND (“cybersecurity” OR “cyber security”) AND (“human-in-the-loop” OR “human in the loop” OR “HITL”) AND (“socio-technical systems”), – (“artificial intelligence” OR “AI” OR “machine learning”) AND (“cybersecurity” OR “cyber security”) AND (“human-AI” OR “human-AI interaction”).

Within the combined platforms, the search returned a total of 265 articles. Then, an unified set of inclusion and quality criteria (Table 3) was applied to refine the selection process (Al-Kadhimi et al., 2023; Zhang et al., 2022), as presented on Table 2. A total of 153 articles remained in the selection process for screening. The scrutiny in selection involved an examination of these studies using their titles and abstracts (Kaur et al., 2023). Articles in dealignment with the topics of this research were excluded, with a remaining set of 102 articles for retrieval.

Table 3 - SLR Inclusion and Quality criteria (own construction)

SLR Inclusion and Quality criteria	
1º.	The article was published after 2015 and in the English language.
2º.	The article was published in a peer-reviewed journal or a conference.
3º.	The topics of the article were in alignment with the topics of this research.
4º.	The article's findings and contributions were relevant to this research.

Following a thorough examination of the full articles and a final assessment of their relevance to this research, it remained a total of 52 articles that were selected through the SLR search strategy. Nevertheless, it was also performed the snowballing approach in the SLR, which refers to using the citations or references within an document to help uncover further relevant studies (Al-Kadhimi et al., 2023), where 6 additional papers were included, bringing the final selection to 58 articles, as outlined in Figure 6.

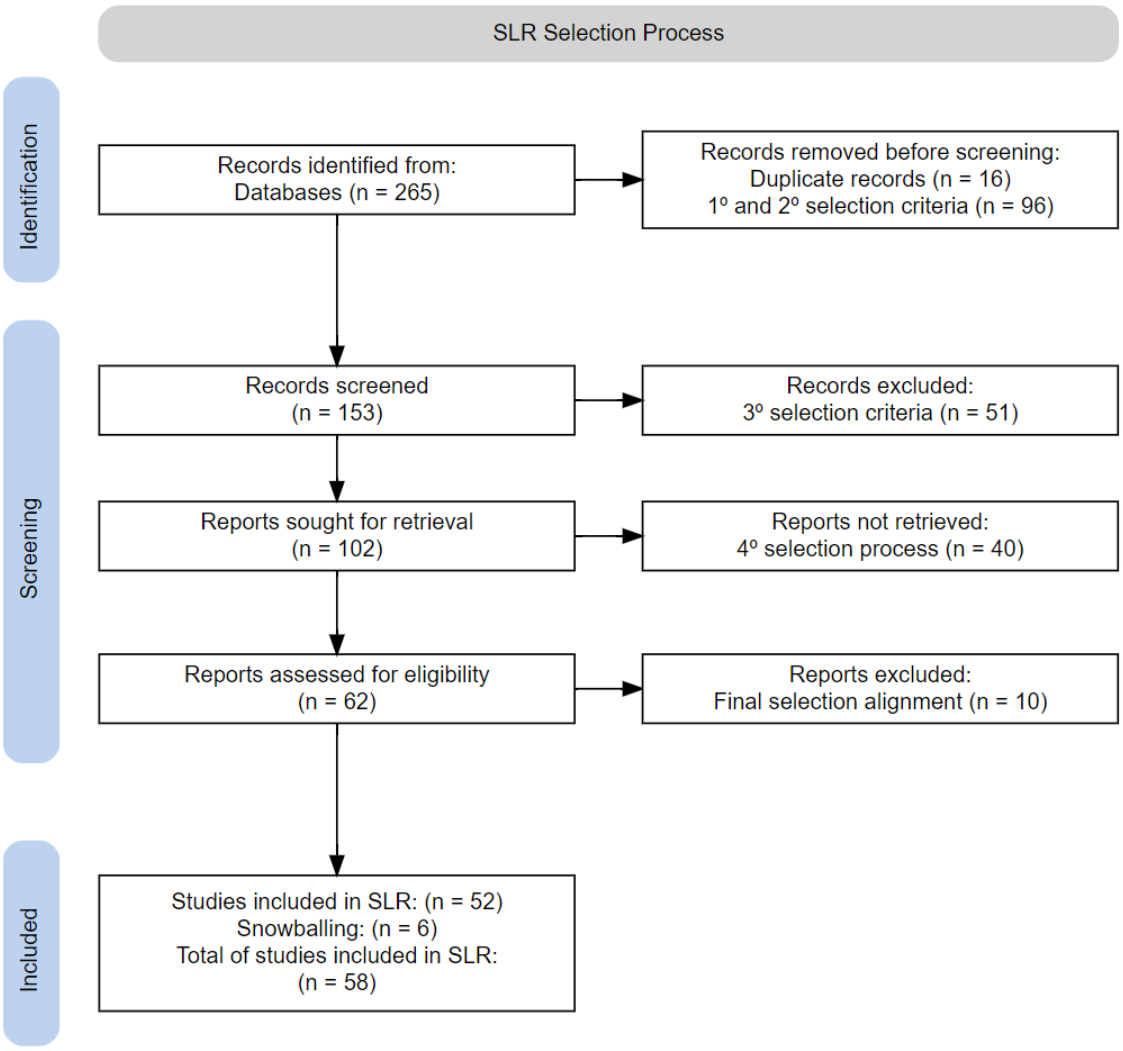

Figure 6 - SLR PRISMA Process – Adapted from (Haddaway et al., 2022; Page et al., 2021)

After adhering to the PRISMA approach from Page et al. (2021) for the Systematic Literature Review (SLR) to gather and analyze the most relevant studies for this research, our aim is to address the research questions, beginning with: "*How can we effectively integrate HITL for AI-driven cybersecurity?"*. To that, the objective is to identify relationships and different concepts through the SLR analysis that can lead to the effective and advantageous incorporation of HITL in AI-driven cybersecurity.

Given the complexity already outlined in this research, and specifically concerning the "effectively" part of this question, the most suitable methodology identified to address that is the Socio-technical IS design science research (STISD) approach (Figure 7) from Carlsson et al. (2011), which states that IS design science research should *"develop abstract knowledge that can be used to design and implement IS initiatives. The knowledge is abstract in the sense that it is not a recipe for designing and implementing a specific IS initiative for a specific organization*" (Carlsson et al., 2011, p. 111).

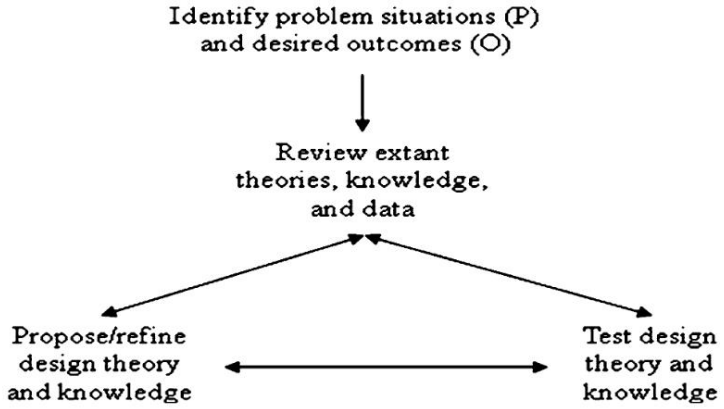

Figure 7 - Socio-technical IS design theory development (Carlsson et al., 2011)

The SLR analysis is performed to gather insights, within the review of extant theories, knowledge, and data from previous studies (Kaur et al., 2023), in parallel with the STISD approach, within the identification of the primary problem situation (Carlsson et al., 2011) as the lack of HITL integration in AI systems for cybersecurity and the primary desired outcome being the advancement in the capabilities of cybersecurity measures through the AI and HITL integration.

To further enhance the research methodology rigor, an instrumental case study methodology (Cousin, 2005) is followed to address our second research question*: "How the integration of HITL for AI-driven cybersecurity contributes to advancing cybersecurity measures in businesses?"*, where to provide a real-life scenario, a case study analysis of the devastating NotPetya cyberattack is performed for understanding cybersecurity challenges and organizational current needs, and serves as a method to simplify the very complex cyberspace threat scenario and theory building of organizational cybersecurity (Ebneyamini & Sadeghi Moghadam, 2018; Gil et al., 2019).

For the instrumental case study research methods process (Figure 8), its purpose is to gather informed conclusions regarding the vulnerabilities that facilitated the attack to emerge, by immersing in the context and reflecting on the findings through a narrative analysis. The focus is on assembling a comprehensive view of the NotPetya cyberattack while meticulously gathering and analyzing all minor details, carefully through triangulation, considering evidence from multiple sources, both archival data interviews and research reviews (Cousin, 2005; Ebneyamini & Sadeghi Moghadam, 2018).

In terms of internal validity, to form a robust chain of evidence, the aim is to ensure that the analysis findings are aligned with the rigorous historical data examination, tabulation, and recombination. As for external validity, the thick description aims to demonstrate the applicability of the analysis findings beyond the specific cyberattack. Considering its similar threat nature, it holds significant potential relevance for other organizations across different business domains (Cousin, 2005; Ebneyamini & Sadeghi Moghadam, 2018).

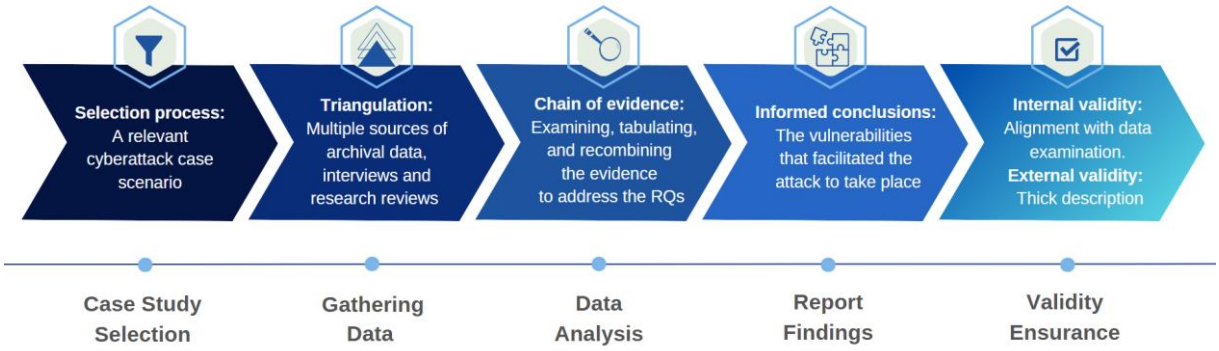

Figure 8 - Instrumental Case Study Process – Based on (Cousin, 2005; Ebneyamini & Sadeghi Moghadam, 2018)

The combination of the rigor and amplitude of the SLR process (Kaur et al., 2023; Page et al., 2021) while also leveraging the depth and specificity of a relevant case study analysis (Cousin, 2005; Ebneyamini & Sadeghi Moghadam, 2018) in the cybersecurity field, aims to integrate both theoretical insights and empirical evidence. Consequently, the socio-technical theory guidance (Bostrom & Heinen, 1977b, 1977a; Carlsson et al., 2011; Taddeo et al., 2023), which is already heavily integrated into this research, serve as the foundational basis for developing a robust new approach. Evaluating its real-world applicability and effectiveness, the proposition of the knowledge (Carlsson et al., 2011) is expected to be presented after the in-depth testing analysis, performed thorough the NotPetya case study scenario.

4. CASE STUDY ANALYSIS

4.1. THE NOTPETYA CYBERATTACK

The NotPetya cyberattack that happened in June 2017, is considered the most impactful and disruptive individual cyberattack in recent history, with its damages surpassing $10 billion (Greenberg, 2018; Ponciano, 2022). Targeting primarily Ukraine, the NotPetya struck in the country hospitals, power providers, airports, banks, ATMs, the national card payment systems and practically all government institutions (Fayi, 2018; Greenberg, 2018).

Credited to the Russian military intelligence, the NotPetya was disguised as a ransomware attack, with its name derived from its similarity to the ransomware known as Petya, which emerged in early 2016 and in the same way demanded victims to pay for a decryption key to regain access to their files. However, NotPetya's bitcoins ransom demands were deceptive, as its code lacked any functionality to provide decryption keys to the victims (Crosignani et al., 2023; Greenberg, 2018; Gupta et al., 2023).

Analyzing the beginning of the NotPetya attack, it had multiple vectors of entry, starting primarily with the Russian military hackers gaining access to the update mechanism of M.E. Doc, a widely used accounting software in Ukraine. Through this intrusion, it was injected malicious code into legitimate software updates, creating a clandestine back door into thousands of PCs worldwide running the M.E. Doc software. This back door access paved the way for the deployment of the NotPetya malware (Crosignani et al., 2023; Greenberg, 2018).

There were also organizations affected through other means, such as social engineering, for example, a victim who opened a PDF file attachment, falsely presented as a job applicant's resume. After the execution of malicious attachments, the ransomware infiltrated the victim's system and encrypted the master file table, which irreversibly encrypted the master boot record of every infected machine, resulting in the effective destruction of these computers and servers. With no need for administrative privileges, the malware spread freely across networks, and while was instated to target only Ukraine, its destructive effects were felt worldwide, affecting different business across diverse sectors (Capano, 2021; Crosignani et al., 2023; Fayi, 2018; Kaminska et al., 2021).

Furthermore, the rapid spread of NotPetya across networks without human interaction was facilitated by a piece of code known as Eternalblue, which exploits a vulnerability found in a specific Windows SMB protocol. Originally developed by the NSA, the EternalBlue was leaked by the hacking group Shadow Brokers in April 2017 (Almashhadani et al., 2022; Crosignani et al., 2023; Greenberg, 2018). Subsequently, Eternalblue was combined with Mimikatz, another tool often utilized by hackers, to allow lateral movement within networks and to facilitate the extraction and exploitation of privileged credentials stored in the memory of the already infected machines. This combination enabled long-term access to the compromised organizational infrastructure and resulted in the rapid infection of a large number of systems

within a short period (Almashhadani et al., 2022; Crosignani et al., 2023; Dargahi et al., 2019; Sharma et al., 2023).

Microsoft had issued a patch for its EternalBlue vulnerability, but some organizations may have delayed or neglected it, as many systems remained vulnerable. The combination of EternalBlue and Mimikatz enabled the infection of unpatched computers and subsequently allowed the extraction of passwords from those systems to infect other patched computers (Fayi, 2018; Greenberg, 2018). Likewise, as the malware had no mechanism to differentiate between targets before installing its payload, it allowed the NotPetya to propagate on the widest scale in the history of cyberattacks (Kaminska et al., 2021).

Following the after attack, using the shipping conglomerate Maersk example, which was one of the most heavily impacted companies with reported losses around $300 million (Crosignani et al., 2023; Greenberg, 2018), they launched a comprehensive response effort. With all its communications channel affected by the attack, their solution was the establishment of an improvised incident response and emergency recovery center in Great Britain, involving hundreds of staff members working nonstop to rebuild their network (Capano, 2021). Maersk had to reinstall around 4,000 servers, 45,000 PCs, and 2,500 applications over ten days (Crosignani et al., 2023).

Maersk also had a setback that occurred during their recovery process, with the absence of a clean backup for the company's domain controllers, which are crucial servers responsible for user authentication and verification. Fortunately, a backup was discovered in the company's Ghana office, as luckily it had been disconnected from the network due to a blackout prior to the attack, preserving a single untainted copy of the domain controller data. However, transferring this backup to the recovery center proved to be a logistical challenge due to Ghana's inadequate public network infrastructure and low bandwidth. In response to the urgent need for a copy, the alternative solution was one employee from Great Britain and another from Ghana had to flew to Nigeria. Once there, the Ghanaian employee personally delivered the physical hard drive to their colleague, who then traveled back to Heathrow Airport with the backup. It took two weeks post-attack to the recovery team deliver most of the clean laptops and computers, wiping existing hard drives, and installing new, uncorrupted copies of the Windows operating system (Capano, 2021; Greenberg, 2018).

The Maersk example stands as just one illustration among the multitude of victims that were extremely affected by the NotPetya attack, that caused significant disruptions and resulted in combined losses of billions of dollars for organizations worldwide (Capano, 2021; Crosignani et al., 2023; Greenberg, 2018). Further losses also came from affected customers that after the cyberattack looked for new alternative suppliers, terminating relations with suppliers affected by the disruptions (Crosignani et al., 2023).

Therefore, the weak spots that enabled NotPetya to propagate extremely fast and destructively not only continue to exist in many organizations (Greenberg, 2018; Jai, 2020;

Ponciano, 2022) but as technology advances becomes more complex than before, where a significant number of contemporary cyber incidents are due to insufficient knowledge and unpreparedness from organizations in identifying and safeguarding against cyberattacks (Chowdhury & Gkioulos, 2021).

Table 4 outlines the key vulnerabilities identified in the instrumental case study analysis, mapping them to the AI-HITL dimensions, vulnerable points, and corresponding cybersecurity measures. Under Proactive Safeguard Control, the intelligence dimension include compromised third-party software, social engineering, weak endpoint security, and unpatched software. for damage control, the resilience dimension addresses lack of network segmentation and inadequate backup procedures. The governance dimension covers access control management and disaster recovery management, with weak passwords, lack of multi-factor authentication, absence of an incident response plan, lack of a disaster recovery plan, communication breakdowns, and regulatory compliance issues. Each vulnerability is discussed to enhance the understanding within the context of the NotPetya case study chain of evidence.

Table 4 - Vulnerable points from the NotPetya Case Study (own construction)

Dimension	Measures	Vulnerable point	Description
Intelligence	*Proactive safeguard control*	**Compromised Third-Party Software**	The lack of code reviews and digital signature verification processes allowed the infection through the M.E. Doc. Software.
		Social Engineering	Employees opened malicious attachments disguised as a job applicant's resume, leading to malware infection.
		Weak Endpoint Security	Weak endpoint security measures, being outdated antivirus software or ineffective intrusion detection systems, as NotPetya easily infected endpoints.
Resilience		**Unpatched Software**	Unpatched versions of Windows SMB protocol enabled NotPetya to spread the malware laterally across other unpatched systems, via the EternalBlue exploit.
	Damage control	**Lack of Network Segmentation**	NotPetya spread laterally across interconnected networks of different organizations because there was no proper network segmentation.
		Inadequate Backup Procedures	Inadequate or outdated backup procedures led to organizations losing critical data during the NotPetya attack, affecting data recovery and restoration of business operations.

Governance	*Access control management*	**Weak Passwords**	Having default and weak passwords on user and administrative credentials, NotPetya gained initial access to systems, facilitating for Mimikatz to extract credentials and escalate privileges, allowing the malware to move laterally across the network.
		Lack of Multi-factor Authentication	With the absence of multi-factor authentication, NotPetya allowed the attackers to easily compromise user accounts.
	Disaster and Recovery management	**Lack of Incident Response Plan and Disaster Recovery Plan**	Not having a predefined incident response plan and a disaster recovery plan, organizations faced confusion and delays in coordinating response efforts and implementing important mitigation measures.
		Communication Breakdowns	Lack of clear and coordinated communication channels negatively impacted the decision-making processes of effective incident response efforts.
		Regulatory Compliance Issues	Following the NotPetya attack, ensuring compliance regulations and notifying relevant authorities added complexity to the response and recovery process.

With the critical primary reminder of the need to worldwide businesses increase cybersecurity preparedness (Crosignani et al., 2023; Greenberg, 2018; Jai, 2020), examining the informed conclusions, which are the vulnerable points from the NotPetya attack, can help to enlighten on how organizational cybersecurity can increase its capabilities, adopt a proactive posture against the ever-evolving cyber threats and advance cybersecurity countermeasures to minimize the potential impact of future attacks.

Following Ebneyamini & Sadeghi Moghadam (2018) approach on case study validity, the findings are aligned with the historical data examination and remain pertinent (Ebneyamini & Sadeghi Moghadam, 2018) to the contemporary cyberspace landscape. Despite the passage of time since the NotPetya attack, the informed conclusions retain significant relevance as a prism for understanding and addressing organizational cybersecurity challenges, covering diverse industries and business types. The proposed AI-HITL dimensions were identified within its respective measures in the vulnerabilities findings, providing not only insights for enhancing cybersecurity measures, but also offering guidance potential, which applies not only to this type of cyberattack but also to a vast range of cyber threats worldwide.

5. RESULTS AND DISCUSSION

5.1. The AI-HITL-Driven Cybersecurity Approach

After the validation of the dimensions of the AI-HITL-driven cybersecurity capabilities, finally, this research suggests a structured representation of the AI-HITL approach and expected contribution for cybersecurity (Figure 9).

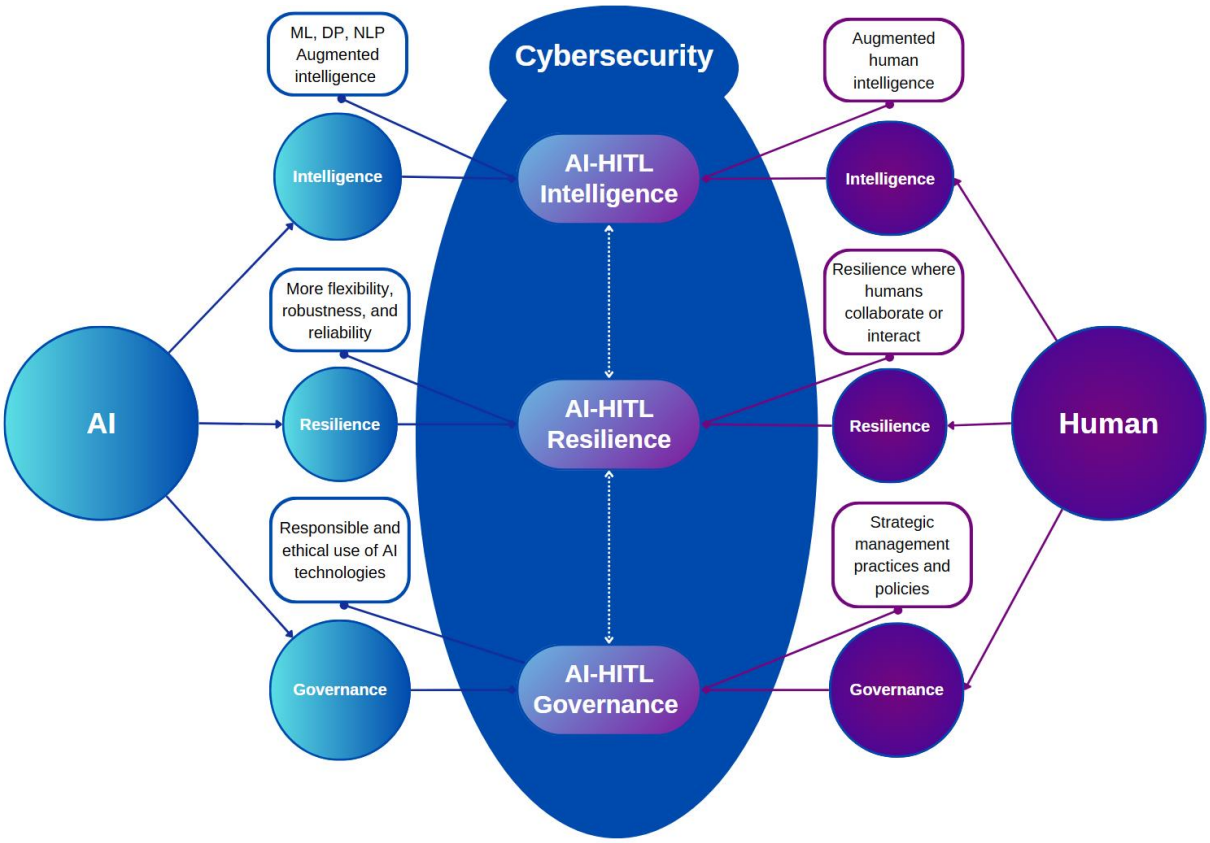

Figure 9 - The AI-HITL in Cybersecurity (own construction)

As illustrated in Figure 9, the AI-HITL approach (AI-HITL) for cybersecurity is composed of the dimensions of AI-HITL in cybersecurity capabilities, represented as AI-HITL Intelligence, AI-HITL Resilience and AI-HITL Governance, thoroughly explored already in this research. In this model, complementarity is still the core idea. Each dimension is interconnected and is mutually influenced in the cybersecurity domain. This AI-HITL dynamic interaction enables AI capabilities (AIC) and HITL capabilities (HITLC) collaboration to happen in a way that alternates or changes over time, facilitating the exchange and complementarity of capabilities in the ever-changing cybersecurity field.

5.2. The AI-HITL-Driven Cybersecurity Core Components

While acknowledging the multitude and complex landscape within the integration of AIC and HITLC, this research endeavors to provide a discussion on the strategic application of the AI-HITL approach, by synthesizing diverse perspectives and discussing possibilities. This strategy not only aims to address current cybersecurity challenges but also to anticipate future advancements in harnessing AIC and HITLC for improving cybersecurity countermeasures.

Therefore, we start by delineating the AI-HITL approach core components. Adapted from Carlsson et al. (2011), Table 5 depicts how the approach is followed, including step-by-step core and additional components, along with their corresponding descriptions (Carlsson et al., 2011).

Table 5 - The AI-HITL Core Components – Adapted from (Carlsson et al., 2011)

Core components	Description
Purpose and Scope	Advancement in the capabilities of cybersecurity countermeasures for organizations through the AI-HITL integration.
Constructs	Cybersecurity capabilities, AI capabilities, HITL capabilities.
Principles (P)	#P1: Harnessing hybrid intelligence increases cybersecurity intelligence capabilities, advancing strategic defense and counterattack. #P2: Combining AI and HITL resilience improves cybersecurity resilience capabilities, enhancing the robustness and reliability of estimations. #P3: Positioning governance as a strategic tool, where socio-technical policies are equally important as algorithm performance ensures efficient, tactical, and responsible cybersecurity measures.
Principles of Implementation (PI)	#PI1: Customize the AI-HITL to the specific scenario by defining strategies that optimize the exchange of capabilities, collaboration, and effectiveness. #PI2: Ensure that the customization increases cybersecurity capabilities, assigning the responsible(s), AI or HITL. We suggest delineating roles and responsibilities accordingly with each measure and respective dimension into Design, Communication, Action, Decision, Control, Monitoring, Evaluation and Continuous Improvement. #PI3: Evaluation of the results needs to be constant to ensure adaptation, sustainability, and continuous improvement. #PI4: Team engagement and awareness are essential for a successful implementation. Training and follow-up sessions need to be provided.
Artifact Mutability	Feedback and suggestions on the AI-HITL approach are crucial for steady improvement.
Expository Instantiation	The design propositions are demonstrated (through Figures 10 to 13, and Table 6) on how the AI-HITL approach can be applied in real-life scenarios.

5.3. The Dimensions of the AI-HITL Approach in Cybersecurity

5.3.1. AI-HITL Intelligence

The AI-HITL intelligence approach aims to combine the potential of AI and humans, for example, having the goal to enhance human decision-making through machine predictions, but keeping the HITL approach, with humans still using its instinct and understanding to make the final decision (Dellermann et al., 2019; Metcalf et al., 2019). As well as, using human intelligence with the purpose of enhance the training of ML models, for example through labelling datasets, guaranteeing that the algorithm aligns with the organization's needs (Dellermann et al., 2019; Grønsund & Aanestad, 2020; Johnson et al., 2022). This approach empowers organizations to foresee and identify anomalies and also use human expertise to provide contextual understanding, correctly address ambiguities and act on situations that were dismissed or wronged by AI systems (Namvar et al., 2023; Zimmermann & Renaud, 2019).

As illustrated on Figure 10, complementarity plays a crucial role for organizational cybersecurity, with AI diverse techniques functioning as a superior observer and even problem solver, as it can consistently monitor network activity, does not have human fatigue and autonomously determines optimal actions, consistently adjusting to its environment (Huang & Rust, 2018; Michael et al., 2023). However, AI also shares vulnerabilities with other computing systems, where unexpected elements can compromise its effectiveness (Michael et al., 2023), and to address that, the effective combination of HITL methods, especially in decision-making, with AI systems distinctive complementary capabilities can lead to the improvement of "intelligent" performance in complex situations (Caldwell et al., 2022; Wu et al., 2022).

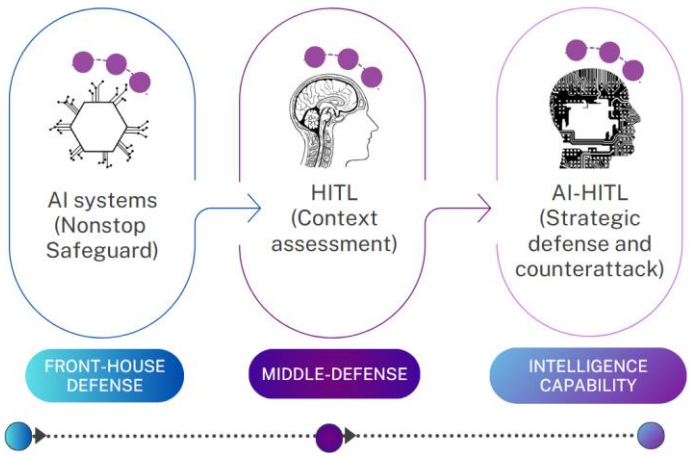

Figure 10 - AI-HITL Intelligence (own construction)

5.3.2. AI-HITL Resilience

In the AI-HITL resilience approach, cybersecurity protective AI technologies ensure safeguard and resilience of systems for organizations, using tamper-evident features that can safeguard

against breaches, non-authorized changes, and sensitive data extraction (Kaur et al., 2023), at the same time that organizations can leverage the collective knowledge, contextual understanding and expertise of multiple experts to enhance the robustness of estimations (Henriques De Gusmão et al., 2018).

Complementarity is also critical, as illustrated on Figure 11, and it can be transferred via the increase of the algorithms capabilities (Fügener et al., 2021; Nunes et al., 2015). By establishing a feedback loop that facilitates the learning process, continuous adjustments and expansion, human input and collaboration improves the performance and resilience of systems by complementing the lack of unique human knowledge and perception. It enhances the algorithm performance, organizational resilience and reflexivity (Fügener et al., 2021; Grønsund & Aanestad, 2020). Likewise, by exchanging AI-HITL capabilities for developing and maintaining resilient measures, cyber resilience is guaranteed both during and after threat events (Blair et al., 2019; Zimmermann & Renaud, 2019).

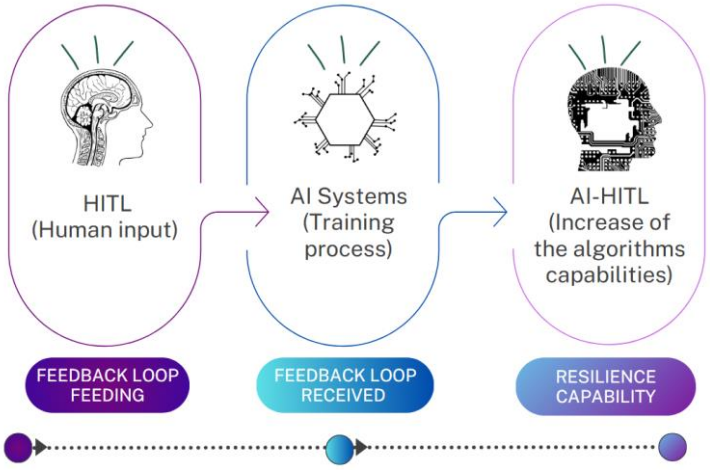

Figure 11 - AI-HITL Resilience (own construction)

5.3.3. AI-HITL Governance

The AI-HITL Governance approach aims to an advanced governance proposition that addresses the transparency, interpretability and explainability challenges when humans are dealing with machine decision-making (Caldwell et al., 2022). While tracking the rationale behind decisions made by explainable ML, DL, and NLP systems, humans are the ultimate decision-makers (Zanzotto, 2019).

Complementarity is achieved through AI predictive analytics, which anticipate and prioritize cybersecurity risks effectively, recommending appropriate measures, and through automated policy enforcement (Kaur et al., 2023), combined with humans acting on compliance requirements, overall business awareness and on the development of strategies that guarantees that the performance of the algorithms aligns with the organization needs (Cram et al., 2017; Grønsund & Aanestad, 2020). Aiming for an advanced cybersecurity governance

that effectively supervise, monitor, review and continuous improve the policies collection for the trustworthiness of AI-HITL utilization (Cram et al., 2017; Kaur et al., 2023; Taeihagh, 2021), as illustrated on Figure 12.

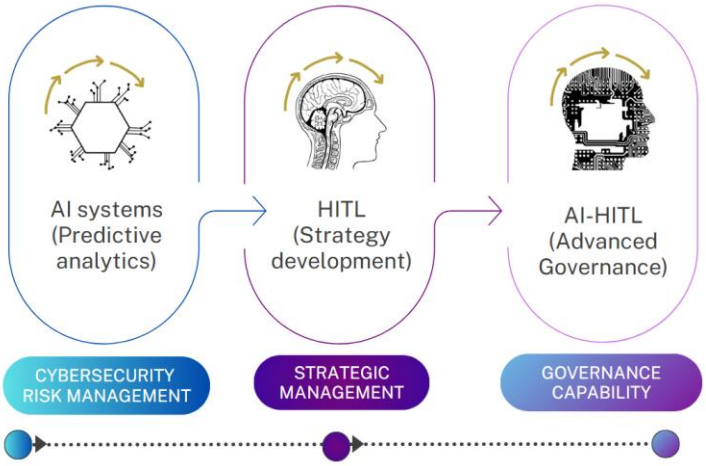

Figure 12 - AI-HITL Governance (own construction)

It is also important to recognize that one single solution approach is not realistic, as the challenge of designing a combined cybersecurity corporate governance is increased by the cyber threat landscape in constant changes. Continuous improvement in the AI-HITL governance policies is essential, where corrections and adaptations are constantly needed in near term periods, allowing organizations to keep up with ever coming challenges, assessing its respective impact and ensuring proper response and recovery strategies (de Bruijn & Janssen, 2017; Domínguez-Dorado et al., 2022; Jalali et al., 2019).

5.4. The AI-HITL in Cybersecurity Teams

Therefore, moving to the exemplification of the real-world applicability and effectiveness of the AI-HITL approach, as already explained its concepts, Figure 13 showcases the approach to cybersecurity blue and red teams.

In the proactive defense team, the AI-HITL Blue team safeguards the system and monitors threats, combining human expertise and AI algorithms, detecting anomalous behavior, poisoning, evasion attacks, and other potential algorithmic attacks, while also addressing the challenges of a top-down decision-making process that could potentially sideline non-managerial roles. In the proactive offensive front, the AI-HITL Red team simulates attacks, continually improving strategies to assess defenses and identify vulnerabilities, while also guaranteeing the processing of large data volumes and up-to-date high-quality datasets, replicating social engineering scenarios and possible interactions, and focusing on continuous learning. The human-AI teamwork integration, between defensive blue and proactive red teams, fosters a culture of learning, growth and shared AI-HITL capabilities. Both human and machine (ACD-based solutions, ML, DL, NLP and other compatible AI techniques), form a

comprehensive strategy for increasing cybersecurity countermeasures (Aoyama et al., 2015; Chindrus & Caruntu, 2023; Feffer et al., 2024; Halisdemir et al., 2022; Titus & Russell, 2023; Vyas et al., 2023).

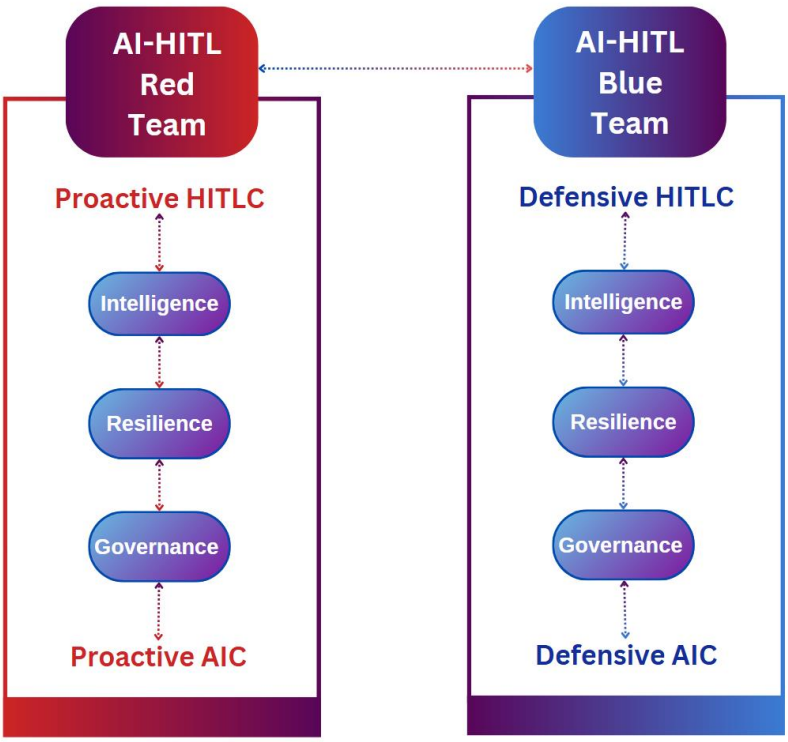

Figure 13 - The AI-HITL approach to cybersecurity teams (own construction)

In the human-AI teamwork, the user point of view is critical, as the success of the integration depends on the willingness of the human team to collaborate with the AI team, with trust growing as automation demonstrates its reliability (Baroni et al., 2022; Caldwell et al., 2022). The criticality of the AI-HITL approach to cybersecurity teams emerges from the current nature of cyberattacks, that shifted from merely causing small and medium disruptions to creating chaos in society, and negatively impacting not only organizations, but also human wellbeing. In fact, it is progressing to get worse, and in this new era, if AI is only partially adopted in cybersecurity, it will weaken cyber defenses by providing a strategic edge to malicious actors who can extensively use AI to initiate new attacks (Sharma et al., 2023; Taddeo, 2019; Teichmann, 2023; Zeadally et al., 2020).

Although not being the focus of this research, the AI-HITL approach also addresses the critical importance of the ethical use of AI and user acceptance (Baroni et al., 2022). The HITL capabilities help in the design of strategies to increase efficiency, transparency, and trust for users. The intelligence and the resilience dimensions, incorporating human knowledge and contextualization, provide a more user-friendly AI-driven cybersecurity experience, where errors and biases in automated threat detection and response are minimized. The Governance dimension acts as a bridge to ensure AIC responsible conduct, to safeguard individual rights, and the proper differentiation between legitimate and illegitimate actors and targets.

That means to guarantee not only secure, but also unbiased, and accountable outcomes to serve the cybersecurity field for organizations, where ethical evaluations in the application stage are needed. It is pivotal for the AI team to articulate its decisions and actions to the human team, as ultimately, humans, not AI, will bear the responsibility for the outcomes of those actions (Ahmad et al., 2023; Hauptman et al., 2023; Taddeo, 2019; Taddeo et al., 2023; Titus & Russell, 2023).

Besides, specially related to human-associated risks, as already discussed, is not advisable to only rely on AI and exclude human involvement. Humans remain essential in-the-loop of enhancements and decision-making processes, for systematic individual reflection on decisions and traceability throughout the entire algorithm process thinking, and to improve overall measures and countermeasures capabilities (Coombs et al., 2021; Michael et al., 2023; Sánchez-García et al., 2023). This potential integration presents a promising opportunity to improve results, strengthen defenses and withstand complex challenges in the cybersecurity domain (Wu et al., 2022; Zhang et al., 2022).

5.5. The AI-HITL Approach to the NotPetya

Furthermore, to demonstrate the validation of the AI-HITL approach, on Table 6, the previous identification of the vulnerable points from the NotPetya case study (Table 4), serves as a real-life cyberattack case scenario to evaluate the AI-HITL practical application. Each vulnerability is addressed within its corresponding measure, and complementarity is achieved through a combined dynamic use of HITLC and AIC, which is customizable to different scenarios. AI capabilities, which involve automation and different AI techniques, reach all three dimensions. Meanwhile, HITL capabilities which involve human expertise, contextual understanding, and strategic management, also cover the three dimensions. With a variety of AI techniques as displayed on Table 1, there is no right capability to use or a certain order to follow, as the interconnected dimensions focus on feedback loops and continuous improvement, to elevate cybersecurity capabilities for the constantly evolving threats in cyberspace.

Revisiting our research questions, in addressing the RQ1: How to effectively integrate HITL for AI-driven cybersecurity? The findings suggest that our proposed AI-HITL approach matches the unexplored opportunity to incorporate HITL into AI-driven cybersecurity. The approach outlines specific dimensions and provides a detailed roadmap for implementation.

As for our RQ2: How does the integration of HITL for AI-driven cybersecurity contribute to advancing cybersecurity measures in businesses? Our findings indicate that the AI-HITL integration significantly enhances cybersecurity countermeasure capabilities. The strategic application finish line is specifically with the successful interconnectedness of the Intelligence, Resilience and Governance dimensions for cybersecurity use cases, offering endless possibilities for future applications. This interconnectedness strengthens existing cybersecurity strategies and creates a flexible and solid foundation for responding to emerging threats.

Table 6 - The AI-HITL approach to the NotPetya (own construction)

D	M	Vulnerability	Description	AI-HITL Approach
Intelligence	*Proactive safeguard control*	***Compromised Third-Party Software***	The lack of code reviews and digital signature verification processes allowed the infection through the M.E. Doc. Software.	**AIC:** nonstop monitoring and analyzing of third-party software for suspicious behavior or unauthorized access. **HITLC:** contextual analysis and decision-making capabilities to differentiate anomalies and respond appropriately.
		Social Engineering	Employees opened malicious attachments disguised as a job applicant's resume, leading to malware infection.	**AIC:** automated email analysis, internal messages content and communication patterns. **HITLC:** expertise in threat awareness and response to identify phishing attempts or suspicious communications.
		Weak Endpoint Security	Weak endpoint security measures, being outdated antivirus software or ineffective intrusion detection systems, as NotPetya easily infected endpoints.	**AIC:** real-time threat detection, and automatic response to security incidents. **HITLC:** design and implementation of effective and personalized pre-defined actions on threats.
Resilience		***Unpatched Software***	Unpatched versions of Windows SMB protocol enabled NotPetya to spread the malware laterally across other unpatched systems, via the EternalBlue exploit.	**AIC:** automated risk assessment analysis and identification of unpatched software. **HITLC:** supervision and prioritization of patch deployment processes.
	Damage control	***Lack of Network Segmentation***	NotPetya spread laterally across interconnected networks of different organizations because there was no proper network segmentation.	**AIC:** segmentation of network traffic based on automated behavior analysis and access control policies. **HITLC:** design and implementation of effective and personalized network segmentation strategies.
		Inadequate Backup Procedures	Inadequate or outdated backup procedures led to organizations losing critical data during the NotPetya attack, affecting data recovery and restoration of business operations.	**AIC:** automation and assurance of regular backup process, such as offline snapshots. Automatic and constant integrity checks to verify data reliability. **HITLC:** management of the backup integrity checks and creation of appropriate recovery plans to mitigate the data damages of possible attacks.

Governance	*Access control management*	***Weak Passwords***	Having default and weak passwords on user and administrative credentials, NotPetya gained initial access to systems, facilitating for Mimikatz to extract credentials and escalate privileges, allowing the malware to move laterally across the network.	**AIC:** automated audits and reinforcement of passwords policies. **HITLC:** design and implementation of appropriate policies and assurance of user awareness, education, and training.
		Lack of Multi-factor Authentication	With the absence of multi-factor authentication, NotPetya allowed the attackers to easily compromise user accounts.	**AIC:** enforcing multi-factor authentication adoption across user accounts. **HITLC:** definition of the most appropriate additional layer of security to prevent unauthorized access even if the stored credentials are compromised.
	Disaster and Recovery management	***Lack of Incident Response Plan and Disaster recovery plan***	Not having a predefined incident response plan and a disaster recovery plan, organizations faced confusion and delays in coordinating response efforts and implementing important mitigation measures.	**AIC:** automated incident detection, analysis, and response workflows. **HITLC:** development and implementation of appropriate incident response plans.
		Communication Breakdowns	Lack of clear and coordinated communication channels negatively impacted the decision-making processes of effective incident response efforts.	**AIC:** automatic alerts of substitute communication channels. **HITLC:** pre-definition of strategies for regular team status updates until full recovery.
		Regulatory Compliance Issues	Following the NotPetya attack, ensuring compliance regulations and notifying relevant authorities added complexity to the response and recovery process.	**AIC:** automation of compliance processes and alerts of issues. **HITLC:** design and implementation of controls and processes to effectively address compliance.

Note. D = Dimension; M = Measures.

It is important to highlight that although the insights gained from the NotPetya cyberattack case study serve as support to attempt to alter reality, demonstrating the effectiveness of the AI-HITL approach, it is displayed in a generalized approach. The validation process is based on a theoretical framework, specifically following the STISD approach proposed by Carlsson et al. (2011), thus, the significant limitation of this research is that the AI-HITL approach awaits validation through real-world implementation and towards saturation (Carlsson et al., 2011).

6. CONCLUSIONS AND FUTURE WORKS

The significance of this research begins with addressing the gap in cybersecurity academia, where studies often overlook the human element or only stress about humans being the problem when discussing artificial intelligence for cybersecurity. Following the systematic literature review, it was found that in the context of the rapid advancements in AI techniques for cybersecurity, humans are frequently suggested to be left out-of-the-loop, as a major proportion of cyberattacks are correlated with human errors and mistakes. However, completely excluding humans also eliminates the unique contributions they bring to cybersecurity countermeasures.

This research also introduces an approach that goes beyond the typical Human-AI collaboration, by incorporating dimensions that directly influence cybersecurity countermeasures within organizations. The development of the proposed AI-HITL approach is unique, as there were no comparable approaches or frameworks at the time of this study. This uniqueness is from the fact that most cybersecurity studies found during the searching process excludes human involvement, confirming the relevance of this research, which on the contrary, integrates the strengths of both humans and AI with an emphasis on continuous improvement.

When this research theoretical and empirical evidence is viewed comprehensively, it confirms that humans are crucial to be kept in-the-loop. The application of the AI-HITL approach to the NotPetya vulnerable points (Table 6) led to the enhancement of cybersecurity countermeasures, evidencing the augmentation of intelligence, improvement of resilience, and fostering of management practices and policies governance by synthesizing human capabilities with AI capabilities. Presenting a promising avenue for future research, it addresses the criticality of leveraging social-technical factors in organizational cybersecurity and it is hoped to inspire further efforts towards fomenting it, as feedback from other researchers and industries specialists implementing the approach will be essential in its refinement. This is critical to ensure the effective approach adaptation in several types of organizations and its continual development proposed nature.

Through the proposed AI-HITL approach, customizing it to meet their specific needs, organizations can create more balanced and effective cybersecurity strategies. Not by chance, complementarity, repeated over this research is the main finding that not only supports, but guides the potential of the AI-HITL approach. Finally, in the future of AI-driven cybersecurity, mastering the AI-HITL integration, specifically the where and how to keep humans in-the-loop, will enable organizations to leverage this duality as an asset rather than a liability, advancing cybersecurity measures to achieve a higher level of security, and raising its preparedness in the ever-evolving cyber threat landscape.

BIBLIOGRAPHICAL REFERENCES

Ahmad, R., Alsmadi, I., Alhamdani, W., & Tawalbeh, L. (2023). Zero-day attack detection: A systematic literature review. *Artificial Intelligence Review*, *56*(10), 10733–10811. https://doi.org/10.1007/s10462-023-10437-z

Al-Kadhimi, A. A., Singh, M. M., & Khalid, M. N. A. (2023). A Systematic Literature Review and a Conceptual Framework Proposition for Advanced Persistent Threats (APT) Detection for Mobile Devices Using Artificial Intelligence Techniques. *Applied Sciences*, *13*(14), Article 14. https://doi.org/10.3390/app13148056

Almashhadani, A. O., Carlin, D., Kaiiali, M., & Sezer, S. (2022). MFMCNS: A multi-feature and multi-classifier network-based system for ransomworm detection. *Computers & Security*, *121*, 102860. https://doi.org/10.1016/j.cose.2022.102860

Al-Saraireh, J., & Masarweh, A. (2022). A novel approach for detecting advanced persistent threats. *Egyptian Informatics Journal*, *23*(4), 45–55. https://doi.org/10.1016/j.eij.2022.06.005

Aoyama, T., Naruoka, H., Koshijima, I., Machii, W., & Seki, K. (2015). Studying resilient cyber incident management from large-scale cyber security training. *2015 10th Asian Control Conference (ASCC)*, 1–4. https://doi.org/10.1109/ASCC.2015.7244713

Astra. (2024, January 16). *160 Cybersecurity Statistics: Updated Report 2024*. Astra. https://www.getastra.com/blog/security-audit/cyber-security-statistics/

Baroni, I., Calegari, G. R., Scandolari, D., & Celino, I. (2022). AI-TAM: A model to investigate user acceptance and collaborative intention in human-in-the-loop AI applications. *Human Computation*, *9*(1), Article 1. https://doi.org/10.15346/hc.v9i1.134

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, *45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Bertia, A., Xavier, S. B., Kathrine, G. J. W., & Palmer, G. M. (2022). A Study about Detecting Ransomware by Using Different Algorithms. *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 1293–1300. https://doi.org/10.1109/ICAAIC53929.2022.9792587

Blair, J. R. S., Hall, A. O., & Sobiesk, E. (2019). Educating Future Multidisciplinary Cybersecurity Teams. *Computer*, *52*(3), 58–66. https://doi.org/10.1109/MC.2018.2884190

Bostrom, R. P., & Heinen, J. S. (1977a). MIS Problems and Failures: A Socio-Technical Perspective. Part I: The Causes. *MIS Quarterly*, *1*(3), 17–32. https://doi.org/10.2307/248710

Bostrom, R. P., & Heinen, J. S. (1977b). MIS Problems and Failures: A Socio-Technical Perspective, Part II: The Application of Socio-Technical Theory. *MIS Quarterly*, *1*(4), 11–28. https://doi.org/10.2307/249019

Caldwell, S., Sweetser, P., O'Donnell, N., Knight, M. J., Aitchison, M., Gedeon, T., Johnson, D., Brereton, M., Gallagher, M., & Conroy, D. (2022). An Agile New Research Framework for Hybrid Human-AI Teaming: Trust, Transparency, and Transferability. *ACM Transactions on Interactive Intelligent Systems*, *12*(3), 1–36. https://doi.org/10.1145/3514257

Capano, D. E. (2021, September 30). *Throwback Attack: How NotPetya Ransomware Took Down Maersk*. Industrial Cybersecurity Pulse. https://www.industrialcybersecuritypulse.com/threats-vulnerabilities/throwback-attack-how-notpetya-accidentally-took-down-global-shipping-giant-maersk/

Carlsson, S. A., Henningsson, S., Hrastinski, S., & Keller, C. (2011). Socio-technical IS design science research: Developing design theory for IS integration management. *Information Systems and E-Business Management*, *9*(1), 109–131. https://doi.org/10.1007/s10257-010-0140-6

Chandra, S., Shirish, A., & Srivastava, S. C. (2022). To Be or Not to Be …Human? Theorizing the Role of Human-Like Competencies in Conversational Artificial Intelligence Agents. *Journal of Management Information Systems*, *39*(4), 969–1005. https://doi.org/10.1080/07421222.2022.2127441

Chindrus, C., & Caruntu, C.-F. (2023). Securing the Network: A Red and Blue Cybersecurity Competition Case Study. *Information*, *14*(11), Article 11. https://doi.org/10.3390/info14110587

Chowdhury, N., & Gkioulos, V. (2021). Cyber security training for critical infrastructure protection: A literature review. *Computer Science Review*, *40*, 100361. https://doi.org/10.1016/j.cosrev.2021.100361

Cohen, I. G., Babic, B., Gerke, S., Xia, Q., Evgeniou, T., & Wertenbroch, K. (2023). How AI can learn from the law: Putting humans in the loop only on appeal. *Npj Digital Medicine*, *6*(1), Article 1. https://doi.org/10.1038/s41746-023-00906-8

Coombs, C., Stacey, P., Kawalek, P., Simeonova, B., Becker, J., Bergener, K., Carvalho, J. Á., Fantinato, M., Garmann-Johnsen, N. F., Grimme, C., Stein, A., & Trautmann, H. (2021). What is it about humanity that we can't give away to intelligent machines? A European perspective. *International Journal of Information Management*, *58*, 102311. https://doi.org/10.1016/j.ijinfomgt.2021.102311

Cousin, G. (2005). Case Study Research. *Journal of Geography in Higher Education*, *29*(3), 421–427. https://doi.org/10.1080/03098260500290967

Cram, W. A., Proudfoot, J. G., & D'Arcy, J. (2017). Organizational information security policies: A review and research framework. *European Journal of Information Systems*, *26*(6), 605–641. https://doi.org/10.1057/s41303-017-0059-9

Crosignani, M., Macchiavelli, M., & Silva, A. F. (2023). Pirates without borders: The propagation of cyberattacks through firms' supply chains. *Journal of Financial Economics*, *147*(2), 432–448. https://doi.org/10.1016/j.jfineco.2022.12.002

Dargahi, T., Dehghantanha, A., Bahrami, P. N., Conti, M., Bianchi, G., & Benedetto, L. (2019). A Cyber-Kill-Chain based taxonomy of crypto-ransomware features. *Journal of Computer Virology and Hacking Techniques*, *15*(4), 277–305. https://doi.org/10.1007/s11416-019-00338-7

de Bruijn, H., & Janssen, M. (2017). Building Cybersecurity Awareness: The need for evidence-based framing strategies. *Government Information Quarterly*, *34*(1), 1–7. https://doi.org/10.1016/j.giq.2017.02.007

Deb, A., Lerman, K., & Ferrara, E. (2018). Predicting Cyber-Events by Leveraging Hacker Sentiment. *Information*, *9*(11), Article 11. https://doi.org/10.3390/info9110280

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, *61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Domínguez-Dorado, M., Carmona-Murillo, J., Cortés-Polo, D., & Rodríguez-Pérez, F. J. (2022). CyberTOMP: A Novel Systematic Framework to Manage Asset-Focused Cybersecurity From Tactical and Operational Levels. *IEEE Access*, *10*, 122454–122485. https://doi.org/10.1109/ACCESS.2022.3223440

Ebneyamini, S., & Sadeghi Moghadam, M. R. (2018). Toward Developing a Framework for Conducting Case Study Research. *International Journal of Qualitative Methods*, *17*(1), 1609406918817954. https://doi.org/10.1177/1609406918817954

ENISA. (2023). *Artificial Intelligence and Cybersecurity Research* (ENISA Research and Innovation Brief). The European Union Agency for Cybersecurity. https://doi.org/10.2824/808362

Fayi, S. (2018). What Petya/NotPetya Ransomware Is and What Its Remidiations Are. In *Information Technology—New Generations* (pp. 93–100). Springer. https://doi.org/10.1007/978-3-319-77028-4_15

Feffer, M., Sinha, A., Lipton, Z. C., & Heidari, H. (2024). Red-Teaming for Generative AI: Silver Bullet or Security Theater? (arXiv:2401.15897). *arXiv*. https://doi.org/10.48550/arXiv.2401.15897

Forescout Vedere Labs. (2024). *2023 Threat Roundup*. https://www.forescout.com/resources/research-report_2023-threat-roundup/

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI. *MIS Quarterly*, *45*, 1527–1556. https://doi.org/10.25300/MISQ/2021/16553

Gil, M., Albert, M., Fons, J., & Pelechano, V. (2019). Designing human-in-the-loop autonomous Cyber-Physical Systems. *International Journal of Human-Computer Studies*, *130*, 21–39. https://doi.org/10.1016/j.ijhcs.2019.04.006

Greenberg, A. (2018). The Untold Story of NotPetya, the Most Devastating Cyberattack in History. *Wired*. https://www.wired.com/story/notpetya-cyberattack-ukraine-russia-code-crashed-the-world/

Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems*, *29*(2), 101614. https://doi.org/10.1016/j.jsis.2020.101614

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access*, *11*, 80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Systematic Reviews*, *18*(2), e1230. https://doi.org/10.1002/cl2.1230

Halisdemir, Maj. E., Karacan, H., Pihelgas, M., Lepik, T., & Cho, S. (2022). Data Quality Problem in AI-Based Network Intrusion Detection Systems Studies and a Solution Proposal. *2022 14th International Conference on Cyber Conflict: Keep Moving! (CyCon)*, *700*, 367–383. https://doi.org/10.23919/CyCon55549.2022.9811014

Hauptman, A. I., Schelble, B. G., McNeese, N. J., & Madathil, K. C. (2023). Adapt and overcome: Perceptions of adaptive autonomous agents for human-AI teaming. *Computers in Human Behavior*, *138*, 107451. https://doi.org/10.1016/j.chb.2022.107451

Henriques De Gusmão, A. P., Mendonça Silva, M., Poleto, T., Camara E Silva, L., & Cabral Seixas Costa, A. P. (2018). Cybersecurity risk analysis model using fault tree analysis and fuzzy decision theory. *International Journal of Information Management*, *43*, 248–260. https://doi.org/10.1016/j.ijinfomgt.2018.08.008

Heyder, T., Passlack, N., & Posegga, O. (2023). Ethical management of human-AI interaction: Theory development review. *The Journal of Strategic Information Systems*, *32*(3), 101772. https://doi.org/10.1016/j.jsis.2023.101772

Huang, M.-H., & Rust, R. T. (2018). Artificial Intelligence in Service. *Journal of Service Research*, *21*(2), 155–172. https://doi.org/10.1177/1094670517752459

Jai, V. (2020). *3 Years After NotPetya, Many Organizations Still in Danger of Similar Attacks*. https://www.darkreading.com/threat-intelligence/3-years-after-notpetya-many-organizations-still-in-danger-of-similar-attacks

Jalali, M. S., Siegel, M., & Madnick, S. (2019). Decision-making and biases in cybersecurity capability development: Evidence from a simulation game experiment. *The Journal of Strategic Information Systems*, *28*(1), 66–82. https://doi.org/10.1016/j.jsis.2018.09.003

Jia, Y., Gu, Z., Du, L., Long, Y., Wang, Y., Li, J., & Zhang, Y. (2023). Artificial intelligence enabled cyber security defense for smart cities: A novel attack detection framework based

on the MDATA model. *Knowledge-Based Systems*, *276*, 110781. https://doi.org/10.1016/j.knosys.2023.110781

Johnson, M., Albizri, A., Harfouche, A., & Fosso-Wamba, S. (2022). Integrating human knowledge into artificial intelligence for complex and ill-structured problems: Informed artificial intelligence. *International Journal of Information Management*, *64*, 102479. https://doi.org/10.1016/j.ijinfomgt.2022.102479

Kaminska, M., Broeders, D., & Cristiano, F. (2021). Limiting Viral Spread: Automated Cyber Operations and the Principles of Distinction and Discrimination in the Grey Zone. *2021 13th International Conference on Cyber Conflict (CyCon)*, 59–72. https://doi.org/10.23919/CyCon51939.2021.9468290

Kane, G. C., Young, A. G., Majchrzak, A., & Ransbotham, S. (2021). Avoiding an Oppressive Future of Machine Learning: A Design Theory for Emancipatory Assistants. *MIS Quarterly*, *45*(1), 371–396. https://doi.org/10.25300/MISQ/2021/1578

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion*, *97*, 101804. https://doi.org/10.1016/j.inffus.2023.101804

Keshk, M., Koroniotis, N., Pham, N., Moustafa, N., Turnbull, B., & Zomaya, A. Y. (2023). An explainable deep learning-enabled intrusion detection framework in IoT networks. *Information Sciences*, *639*, 119000. https://doi.org/10.1016/j.ins.2023.119000

Malatji, M., Marnewick, A., & von Solms, S. (2020). Validation of a socio-technical management process for optimising cybersecurity practices. *Computers & Security*, *95*, 101846. https://doi.org/10.1016/j.cose.2020.101846

Metcalf, L., Askay, D. A., & Rosenberg, L. B. (2019). Keeping Humans in the Loop: Pooling Knowledge through Artificial Swarm Intelligence to Improve Business Decision Making. *California Management Review*, *61*(4), 84–109. https://doi.org/10.1177/0008125619862256

Michael, K., Abbas, R., & Roussos, G. (2023). AI in Cybersecurity: The Paradox. *IEEE Transactions on Technology and Society*, *4*(2), 104–109. https://doi.org/10.1109/TTS.2023.3280109

Minkkinen, M., & Mäntymäki, M. (2023). Discerning Between the "Easy" and "Hard" Problems of AI Governance. *IEEE Transactions on Technology and Society*, *4*(2), 188–194. https://doi.org/10.1109/TTS.2023.3267382

Namvar, M., Intezari, A., Akhlaghpour, S., & Brienza, J. P. (2023). Beyond effective use: Integrating wise reasoning in machine learning development. *International Journal of Information Management*, *69*, 102566. https://doi.org/10.1016/j.ijinfomgt.2022.102566

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (Report/Study NIST AI 100-1; p. NIST AI 100-1). *National Institute of Standards and Technology*. https://doi.org/10.6028/NIST.AI.100-1

National Institute of Standards and Technology. (2024). The NIST Cybersecurity Framework (CSF) 2.0 (NIST CSWP 29; p. NIST CSWP 29). *National Institute of Standards and Technology.* https://doi.org/10.6028/NIST.CSWP.29

Neupane, S., Ables, J., Anderson, W., Mittal, S., Rahimi, S., Banicescu, I., & Seale, M. (2022). Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities. *IEEE Access*, *10*, 112392–112415. https://doi.org/10.1109/ACCESS.2022.3216617

Nunes, D., Zhang, P., & Sá Silva, J. (2015). A Survey on Human-in-the-Loop Applications Towards an Internet of All. *IEEE Communications Surveys & Tutorials*, *17*, 1–1. https://doi.org/10.1109/COMST.2015.2398816

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, *10*(1), 89. https://doi.org/10.1186/s13643-021-01626-4

Ponciano, J. (2022). 'Extremely Destructive' Russian Cyberattacks Could Cost U.S. Billions Of Dollars In Economic Damage, Goldman Warns. *Forbes*. https://www.forbes.com/sites/jonathanponciano/2022/03/07/extremely-destructive-russian-cyberattacks-could-cost-us-billions-of-dollars-in-economic-damage-goldman-warns/

Rana, M. U., Ellahi, O., Alam, M., Webber, J. L., Mehbodniya, A., & Khan, S. (2022). Offensive Security: Cyber Threat Intelligence Enrichment With Counterintelligence and Counterattack. *IEEE Access*, *10*, 108760–108774. https://doi.org/10.1109/ACCESS.2022.3213644

Saha, S., Gan, Z., Cheng, L., Gao, J., Kafka, O. L., Xie, X., Li, H., Tajdari, M., Kim, H. A., & Liu, W. K. (2021). Hierarchical Deep Learning Neural Network (HiDeNN): An artificial intelligence (AI) framework for computational science and engineering. *Computer Methods in Applied Mechanics and Engineering*, *373*, 113452. https://doi.org/10.1016/j.cma.2020.113452

Sánchez-García, I. D., Feliu Gilabert, T. S., & Calvo-Manzano, J. A. (2023). Countermeasures and their taxonomies for risk treatment in cybersecurity: A systematic mapping review. *Computers & Security*, *128*, 103170. https://doi.org/10.1016/j.cose.2023.103170

Sharma, A., Gupta, B. B., Singh, A. K., & Saraswat, V. K. (2023). Advanced Persistent Threats (APT): Evolution, anatomy, attribution and countermeasures. *Journal of Ambient Intelligence and Humanized Computing*, *14*(7), 9355–9381. https://doi.org/10.1007/s12652-023-04603-y

Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., & Xu, M. (2020). A Survey on Machine Learning Techniques for Cyber Security in the Last Decade. *IEEE Access*, *8*, 222310–222354. https://doi.org/10.1109/ACCESS.2020.3041951

Taddeo, M. (2019). Three Ethical Challenges of Applications of Artificial Intelligence in Cybersecurity. *Minds and Machines*, *29*(2), 187–191. https://doi.org/10.1007/s11023-019-09504-8

Taddeo, M., Jones, P., Abbas, R., Vogel, K., & Michael, K. (2023). Socio-Technical Ecosystem Considerations: An Emergent Research Agenda for AI in Cybersecurity. *IEEE Transactions on Technology and Society*, *4*(2), 112–118. https://doi.org/10.1109/TTS.2023.3278908

Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society*, *40*(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377

Teichmann, F. (2023). Ransomware attacks in the context of generative artificial intelligence—An experimental study. *International Cybersecurity Law Review*, *4*(4), 399–414. https://doi.org/10.1365/s43439-023-00094-x

Titus, A. J., & Russell, A. H. (2023). The Promise and Peril of Artificial Intelligence—Violet Teaming Offers a Balanced Path Forward (arXiv:2308.14253; Version 1). *arXiv*. https://doi.org/10.48550/arXiv.2308.14253

Vyas, S., Hannay, J., Bolton, A., & Burnap, P. P. (2023). Automated Cyber Defence: A Review (arXiv:2303.04926). *arXiv*. https://doi.org/10.48550/arXiv.2303.04926

Wiafe, I., Koranteng, F., Obeng, E., Assyne, N., Wiafe, A., & Gulliver, S. (2020). Artificial Intelligence for Cybersecurity: A Systematic Mapping of Literature. *IEEE Access*, *8*, 1–1. https://doi.org/10.1109/ACCESS.2020.3013145

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364–381. https://doi.org/10.1016/j.future.2022.05.014

Zanzotto, F. M. (2019). Viewpoint: Human-in-the-loop Artificial Intelligence. *Journal of Artificial Intelligence Research*, *64*, 243–252. https://doi.org/10.1613/jair.1.11345

Zeadally, S., Adi, E., Baig, Z., & Khan, I. A. (2020). Harnessing artificial intelligence capabilities to improve cybersecurity. *IEEE Access*, *8*, 23817–23837. Scopus. https://doi.org/10.1109/ACCESS.2020.2968045

Zhang, Z., Ning, H., Shi, F., Farha, F., Xu, Y., Xu, J., Zhang, F., & Choo, K.-K. R. (2022). Artificial intelligence in cyber security: Research advances, challenges, and opportunities. *Artificial Intelligence Review*, *55*(2), 1029–1053. https://doi.org/10.1007/s10462-021-09976-0

Zimmermann, V., & Renaud, K. (2019). Moving from a 'human-as-problem" to a 'human-as-solution" cybersecurity mindset. *International Journal of Human-Computer Studies*, *131*, 169–187. https://doi.org/10.1016/j.ijhcs.2019.05.005