NOVA IMS Information Management School

MGI

Master Degree Program in
Information Management

The Dark Side of Artificial Intelligence

AI-Driven Cyber Attacks

Igor Marcelo de Sá

Master Thesis

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

The Dark Side of Artificial Intelligence

AI-Driven Cyber Attacks

by

Igor Marcelo de Sá

Master Thesis presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence

Supervised by

Mijail Naranjo Zolotov, PhD

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.

[Lisbon, July 13th 2024]

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to NOVA IMS, which has been my second home for the past five years.

I would also like to thank professor Mijail Zolotov for his guidance and support in developing this paper.

Finally, a huge thank you to my family and friends who have supported me throughout this journey.

ABSTRACT

The continuous evolution of Artificial Intelligence (AI) has directly impacted the growth of innovation and automation. Unfortunately, this evolution also brings negative effects due to malicious use. Threat actors are constantly changing and improving their attack strategies, emphasizing AI-driven techniques in the attack process. These techniques can be used in conjunction with conventional methods to cause greater damage. This study conducts a systematic literature review to explore existing articles, reports, and empirical studies on AI-based cyber attacks. Out of 104 identified studies, 15 were selected based on relevance and quality criteria. Despite several studies on AI and security, researchers have not sufficiently summarized AI-based cyber attacks to understand adversaries' actions and develop proper defenses. The review aims to identify imminent risks, complex challenges, and potential future scenarios regarding AI-driven cyber attacks. Additionally, it investigates methods for mitigation and countermeasures to defend against such attacks. This comprehensive approach fills the gap in existing literature, providing a deeper understanding of AI-powered cyber threats and offering a framework for future defense strategies.

KEYWORDS

Artificial Intelligence; Cybersecurity; Cyber Attacks; AI-Driven Cyber Attacks; Cyber Defense.

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
APTs	Advanced Persistent Training
IoT	Internet of Things

1. INTRODUCTION

The integration of AI into various aspects of modern life has brought unparalleled advancements, yet it also poses a pressing challenge: the rise of AI-driven cyber attacks. These attacks take advantage of AI's capabilities for malicious intent, causing disruptions in critical systems, compromising data security, and undermining privacy. The alarming pace at which AI technology evolves amplifies the sophistication of cyber threats, posing a significant concern for individuals, organizations, and governments globally (Guembe et al., 2022; Velasco, 2022; Werthner et al., 2022).

The significance of addressing AI-driven cyber attacks is underscored by alarming statistics from prominent global institutions. According to the World Economic Forum's Global Risks Report 2021, cyber attacks continue to rank among the top global risks in terms of likelihood and impact. Notably, the report highlights the exponential rise in AI-driven attacks, which are becoming increasingly sophisticated and challenging to detect or mitigate effectively. The OECD's Digital Economy Outlook further substantiates this concern, revealing a staggering 50% surge in cyber incidents over recent years, where AI augmentation plays a pivotal role in amplifying the scale and intricacy of these threats. Furthermore, studies by the United Nations and Eurostat indicate that businesses across various sectors are experiencing a surge in AI-enabled cyber intrusions, causing substantial economic losses, jeopardizing data integrity, and eroding public trust in digital systems. These statistics unequivocally emphasize the critical need to comprehensively address the burgeoning challenges posed by AI-powered cyber threats before they inflict irreparable damage on global economies, societal trust, and individual privacy.

Despite growing awareness, the literature lacks comprehensive insights into the evolving landscape of AI-driven cyber threats. Current studies often explore either the emerging threat of AI-Driven cyber attacks, challenges and opportunities (Blauth et al., 2022; Comiter, 2019; Hassan & Wasim, 2023), or detecting and mitigating these attacks (Abdullahi et al., 2022; Beg et al., 2023). This study aims to fill this gap by conducting a complete study, merging the existing studies mentioned above, investigating the sophisticated techniques and implications of AI-powered cyber attacks , as well as what the future holds in terms of cyber attacks and also a framework/methodologies to detect future threats and mitigate them, exploring uncharted territory in cybersecurity research.

This study will be conducted throughout a systematic literature review, merging existing studies regarding the emerging threats of AI-Driven cyber attacks, the methods to detect future attacks (framework), as well as mitigations and countermeasures to defense against these attacks (Ansari et al., 2022; Bécue et al., 2021; Blauth et al., 2022; Catania et al., 2018; Comiter, 2019; Guembe et al., 2022; Hassan & Wasim, 2023; Kaloudi & Li, 2021; Li, 2018; Velasco, 2022; Yamin et al., 2021). The goal is to conduct a complete article regarding all of

the aspects mentioned above, due to the fact that most studies only explore each aspect individually.

Based on the research done until now, the study identified a paradigm shift in cyber attack methodologies, showcasing a transition towards AI-Driven tactics that exploit vulnerabilities in existing security frameworks. These attacks leverage AI's adaptability and learning capabilities to evade traditional defense mechanisms, posing unprecedented challenges in threat detection and mitigation. Notably, the study revealed the potential ramifications of such attacks, including data breaches, service disruptions, and compromised integrity of AI-dependent decision-making processes. Furthermore, it emphasized the urgency for adaptive cybersecurity measures that integrate AI-based defense strategies to counteract the evolving sophistication of AI-Driven threats. Overall, these findings underscore the imperative for collaborative efforts among stakeholders to fortify cyber defenses, innovate resilient technologies, and formulate robust policies aimed at safeguarding against the escalating menace of AI-driven cyber attacks.

This research makes significant contributions to both academic understanding and practical implications in the realm of AI-Driven cyber threats. Firstly, it advances scholarly knowledge by providing an in-depth exploration of the evolving landscape of AI-Driven cyber attacks, delving into the intricacies of attack methodologies and their potential consequences. The study contributes to theoretical frameworks in cybersecurity, expanding our comprehension of adversarial machine learning and the intersection of AI and cyber threats. Additionally, the research offers actionable insights for practitioners and policymakers by delineating effective strategies for mitigating AI-driven cyber risks. It outlines adaptive cybersecurity measures that integrate AI-based defense mechanisms, providing a roadmap for organizations and governments to fortify their digital infrastructures. Furthermore, this study serves as a foundational resource for future research endeavors, creating a basis for further exploration into AI ethics, regulation, and the development of secure AI technologies. By addressing the existing gap in the literature and offering tangible contributions, this research aims to empower stakeholders with the knowledge and tools necessary to navigate the dark side of artificial intelligence and foster a more secure and resilient digital future.

2. THEORETICAL BACKGROUND

2.1. ARTIFICIAL INTELLIGENCE

Artificial Intelligence (AI) has undergone substantial evolution since its beginnings, driven by advances in machine learning, neural networks, and data processing capabilities (Goodfellow et al., 2016). AI systems are designed to simulate human intelligence processes, such as learning, reasoning, and self-correction, and have found applications across a wide range of domains, including healthcare, finance, transportation, and cybersecurity (Russell et al., 2022). The ability of AI to process vast amounts of data at unprecedented speeds and derive insights makes it a valuable tool for enhancing efficiency and decision-making processes in both the public and private sectors (Zuiderwijk et al., 2021).

Over the decades, AI has transitioned from simple rule-based systems to more complex algorithms capable of deep learning and autonomous decision-making (LeCun et al., 2015). Key milestones in AI development include the introduction of machine learning algorithms that enable systems to learn from data, and deep learning techniques that utilize multi-layered neural networks to model complex patterns and behaviors (Goodfellow et al., 2016).Modern AI systems are characterized by their ability to perform tasks such as natural language processing, image recognition, and predictive analytics (Russell et al., 2022). These capabilities are powered by sophisticated algorithms that can analyze large datasets to identify trends and make predictions (Halevy et al., 2009). The integration of AI with other emerging technologies such as the Internet of Things (IoT), blockchain, and cloud computing further amplifies its potential, enabling the creation of smart environments and autonomous systems (Zuiderwijk et al., 2021).

The widespread adoption of AI brings significant benefits, including increased efficiency, improved accuracy in decision-making, and the ability to automate routine tasks(Russell et al., 2022). However, it also raises several challenges and concerns. One of the primary issues is the ethical use of AI, particularly regarding privacy, bias, and transparency (Eubanks, 2018). Ensuring that AI systems are fair and unbiased requires rigorous testing and validation, as well as the implementation of ethical guidelines and standards (Floridi et al., 2018). Moreover, the reliance on AI systems introduces new vulnerabilities, particularly in the context of cybersecurity. AI-driven applications are susceptible to adversarial attacks, where malicious actors manipulate input data to deceive AI models (Goodfellow et al., 2016). This highlights the need for robust security measures to protect AI systems from exploitation and ensure their safe deployment (Russell et al., 2022).

2.2. CYBER ATTACKS

Cyber attacks are deliberate attempts by individuals or organizations to breach the information systems of another entity. These attacks aim to steal, alter, or destroy data, disrupt services, or exploit system vulnerabilities for malicious purposes (Anderson et al., 2013). Cyber attacks have evolved significantly over the years, becoming more sophisticated and targeted, posing severe threats to national security, economic stability, and personal privacy (Kshetri, 2018).

Cyber attacks can be classified into various types, including phishing, malware, ransomware, and Distributed Denial of Service (DDoS) attacks. Each type employs different techniques to achieve its objective. For instance, phishing attacks deceive individuals into providing sensitive information by pretending to be a trustworthy entity (Hong, 2012). Malware, on the other hand, involves malicious software designed to damage or gain unauthorized access to systems (Egele et al., 2008). Advanced Persistent Threats (APTs) represent a more sophisticated form of cyber attack, often orchestrated by nation-states or organized crime groups. APTs involve prolonged and targeted efforts to infiltrate and maintain a presence within a network, enabling attackers to gather intelligence or cause significant damage over time (Tankard, 2011). The use of zero-day exploits is also a common technique in sophisticated cyber attacks, which take advantage of previously unknown vulnerabilities, being highly valuable to attackers because they can bypass existing security measures, making them a potent tool in sophisticated cyber attacks (Bilge & Dumitraş, 2012).

The impact of cyber attacks can be devastating, leading to financial losses, reputational damage, and operational disruptions (Kshetri, 2018). High-profile breaches such as the Equifax data breach and the WannaCry ransomware attack underscore the critical need for robust cybersecurity measures (Anderson et al., 2013). Defending against cyber attacks requires a multi-faceted approach that includes technological solutions, organizational policies, and user awareness (NIST, 2018). Key defense mechanisms involve the deployment of firewalls, intrusion detection systems (IDS), and encryption technologies(Stallings, 2017). Regular patching and updates of software and systems are essential to close security gaps (NIST, 2018). Additionally, cybersecurity frameworks and standards, such as the NIST Cybersecurity Framework, provide guidelines for managing and reducing cyber risks (NIST, 2018). Organizations must also invest in training and awareness programs to equip employees with the knowledge to recognize and respond to cyber threats (*SANS Institute*, 2020).

2.3. AI-Driven Cyber Attacks

AI-driven cyber attacks represent a new frontier in cybersecurity threats, leveraging the capabilities of AI to enhance the effectiveness and efficiency of malicious activities (Brundage et al., 2018). These attacks utilize AI to automate tasks, evade detection, and adapt to countermeasures, making them particularly challenging to defend against (Goodfellow et al., 2016).

AI-driven cyber attacks use several sophisticated mechanisms and strategies. Machine learning algorithms can be used to analyze large datasets and identify patterns that indicate potential vulnerabilities (Papernot et al., 2016). AI can also be used to automate the generation of phishing emails or malware, increasing the scale and speed of attacks (Brundage et al., 2018). In some cases, attackers deploy AI to create deepfakes (realistic but fake digital content) that can be used to deceive individuals or systems (Citron & Chesney, 2019). One notable strategy is the use of adversarial machine learning, where attackers intentionally introduce misleading data to corrupt the training process of AI models. This can result in AI systems making incorrect decisions or becoming ineffective (Goodfellow et al., 2016). Additionally, AI-driven social engineering attacks can exploit psychological manipulation techniques to deceive individuals into revealing sensitive information or performing actions that compromise security (Brundage et al., 2018).

The future of AI-driven cyber attacks is marked by increasing sophistication and autonomy. As AI technology continues to evolve, attackers are likely to develop more advanced techniques that can bypass traditional security measures (Brundage et al., 2018). The integration of AI with other technologies, such as IoT and 5G, expands the attack surface, creating new vulnerabilities and opportunities for exploitation (Floridi et al., 2018). One of the significant challenges in addressing AI-driven cyber attacks is the dynamic and adaptive nature of AI. Traditional cybersecurity measures may be insufficient to counter the rapid evolution of AI-driven threats. This necessitates the development of AI-based defense mechanisms that can anticipate and respond to emerging threats in real-time (Russell et al., 2022). Furthermore, the ethical and regulatory implications of AI in cybersecurity must be considered to ensure the responsible use of AI technologies (Floridi et al., 2018).

Effective mitigation of AI-driven cyber attacks requires a comprehensive approach that combines technological, organizational, and regulatory measures. AI-based cybersecurity tools can enhance threat detection and response capabilities by analyzing network traffic, identifying anomalies, and predicting potential attacks (Brundage et al., 2018). These tools must be continuously updated and trained to keep pace with evolving threats (Papernot et al., 2016). Organizations should adopt a proactive approach by conducting regular security assessments, implementing robust access controls, and fostering a culture of cybersecurity awareness (NIST, 2018). Collaboration between industry, universities, and government is essential to develop standards and frameworks that address the unique challenges posed by AI-driven cyber threats (Russell et al., 2022). In conclusion, while AI offers significant benefits

in various domains, its potential misuse in cyber attacks presents a critical challenge. Addressing this threat requires innovative solutions, continuous vigilance, and collaborative efforts to safeguard against the next generation of cyber threats (Brundage et al., 2018).

3. METHODOLOGY

This section describes the approach used in order to conduct a systematic literature review, which was the PRISMA methodology. Therefore, the next chapters describe the stages associated with conducting this review, which are: **1)** Identification of research, **2)** Selection of primary studies, **3)** Study quality assessment, **4)** Data extraction & monitoring and **5)** Data Synthesis.

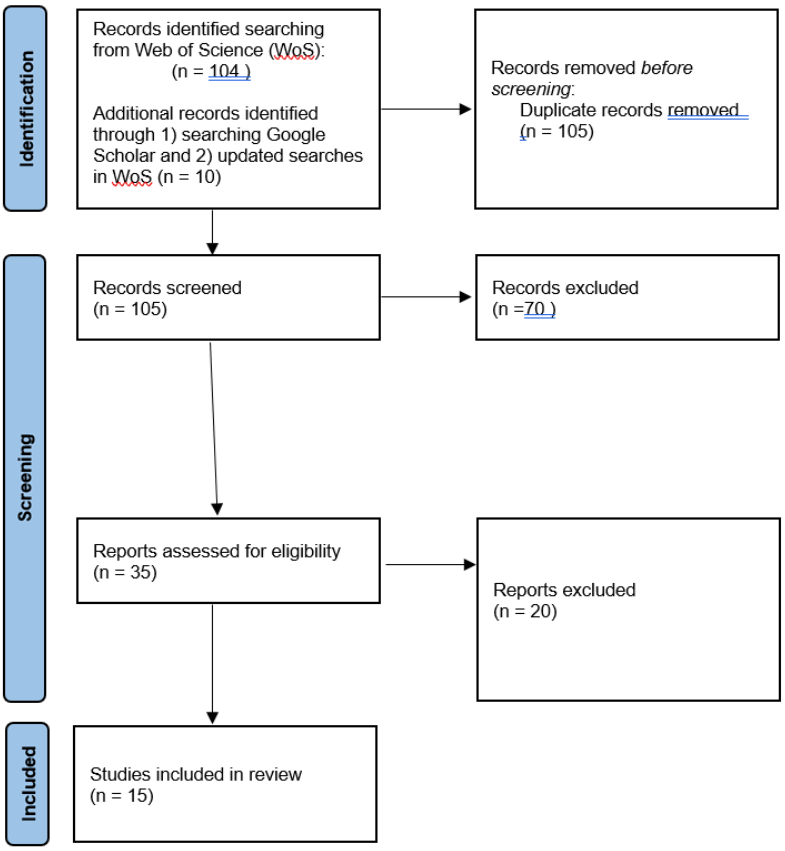

Figure 1 – Study selection, assessment, and inclusion (presented using the PRISMA flow diagram)

3.1. STAGE 1: IDENTIFICATION OF RESEARCH

The first step was to determine the objectives and research questions that shaped our literature review, followed by gathering as much studies related to the research questions as possible without any bias regarding the search strategy (Kitchenham, 2004). The established objectives are the following: **RO1)** contextualize the identified research within the existing knowledge, **RO2)** acquire useful research strategies other journals have used to study AI (and more precisely AI-Driven cyber attacks), and **RO3)** gain valuable perspectives on the implications of the use of AI in cyber attacks.

3.1.1. (RO1): Contextualize the identified research within the existing knowledge

In order to accomplish the first objective, we had to develop the following research questions:

(RQ1). What is the context in which the subject "AI-Driven cyber attacks" have been investigated previously by researchers? (e.g., emerging threats, methods to detect future attacks, mitigations and countermeasures)?

(RQ2). Are the contexts mentioned in above studied individually or is it a complete study of these attacks?

(RQ3). What are the theories and models used in studies regarding AI-Driven Cyber Attacks?

3.1.2. (RO2): Research strategies other journals have used to study AI

In order to accomplish the second objective, the following research question was developed:

(**RQ4).** What was the research approached and methods used in previous studies related to the use of AI in cyber attacks?

3.1.3. (RO3): Implications of the use of AI in cyber attacks

In order to accomplish the third objective, the following research question was developed:

(RQ5). Who are the main entities affected by the use of AI in cyber attacks (direct or indirect)?

(RQ6). What are the main challenges caused by the use of AI in cyber attacks?

(RQ7). Which countermeasures and mitigations are able to defend against AI-Driven cyber attacks?

Two sources were used in order to identify the scientific studies regarding the emerging threats of AI-Driven cyber attacks.The initial search was conducted using Google Schoolar, and then complemented with Web of Science.

3.2. STAGE 2: SELECTION OF PRIMARY STUDIES

The second step consists of the conducted study selection criteria, such as inclusion and exclusion criteria, with the purpose of narrowing the scientific papers to the ones who provide direct evidence about the research question (Kitchenham, 2004). For Web of Science, the search was limited to five categories disciplines: Computer Science Information Systems, Computer Science Theory Methods, Computer Science Artificial Intelligence, Computer Science Sofware Engineering, and Computer Science Hardware Architecture. The research areas were limited to: Computer Science, Engineering. With this, we excluded disciplines, such as: Energy Fuels, Photographic Technology , Telecommunications, Operations Research Management Science, which were considered to not dive into the subject of AI-Driven Cyber Attacks. Also, we limited the results to articles published between 2020-2024 and written in English. Furthermore, we limited our search based on the search terms included in Table 1.

Table 1 - Searched terms used for literature review

Databases	Search terms in the title/keywords
Web of Science, Google Scholar	"Artificial Intelligence" OR "Cybersecurity" OR "Cyber Attacks" OR "AI-Driven cyber attacks" OR "Cyber defense" OR "Malicious use of Artificial Intelligence".

3.3. STAGE 3: STUDY QUALITY ASSESSMENT

The next step is focused on assessing the relevance and quality of the selected studies. This was done firstly by analyzing each of the 104 identified study's title and abstract using the following criteria:

(SQ1). Studies where the use of AI played a significant role in the research questions and objectives were included, whereas studies where the influence of AI was minor were excluded.

(SQ2). AI-driven cyber attacks should be the main priority of the study. Therefore, if the study did not address the use of AI in the context of cyber attacks it was excluded in this phase.

(SQ3). The inclusion of the consequences from the utilization of AI in cyber attacks should be a central topic, examining the ramifications and implications of AI-driven cyber attacks. The studies that did not adress the consequences of the use of AI in cyber attakcs were excluded.

Therefore, based on the criterias metioned above, this process lead to 70 studies being exluded, meaning that there were 35 studies left .

Moreover, the relevance and quality of the remaining 35 studies were assessed by reading the full articles. Each study was independently assessed using quality dimensions derived from (Bano & Zowghi, 2015; Batini et al., 2009):

Accuracy: The study clearly outlines its objectives, providing a detailed description of the data collection methods. Key statements are supported by appropriate references.

Consistency: The study's design aligns well with its research objectives, effectively answering the research questions or achieving the research goals.

Completeness: The research approach is comprehensively detailed.

Timeliness: The study has been published within the last 4 years.

This led to 20 studies being excluded at this stage for reasons including not meeting quality criteria, being opinion-based without a particular research approach, insufficient focus on AI-driven cyber attacks, or being short poster descriptions. This left a final selection of 15 studies.

3.4. STAGE 4: DATA EXTRACTION & MONITORING

To extract data from our literature review, we used a spreadsheet to record the metadata for each of the selected studies. Table 5 depicts the metadata we collected about the 15 selected studies, including descriptive information, approach-related information, quality-related information and AI-driven cyber attacks related information.

3.5. STAGE 5: DATA SYNTHESIS

The final stage of our study involved synthesizing the extracted data to draw meaningful conclusions. This process included integrating findings from various studies to identify common themes and insights, analyzing the synthesized data to comprehend the broader implications of AI-driven cyber attacks, and organizing the results in a coherent manner to provide a comprehensive overview of the topic. By combining data from different sources, we were able to identify patterns and draw conclusions that reflect the current state of research in the field. This synthesis not only highlights the key findings but also underscores the significance of AI-driven cyber attacks, offering a holistic view of the challenges and opportunities within this evolving landscape.

4. RESULTS

This section contains the results obtained from the analysis of the selected research articles, with emphasis on AI-Driven Cyber Attacks. Therefore, the following sections present the results from our descriptive analysis, approach analysis, quality analysis, and content analysis.

4.1. DESCRIPTIVE ANALYSIS

The first section of the analysis contains the study of the selected studies objectives (Table…), the journals and conferences where these studies appeared, their years of publication, and the databases used to locate them. Initially, the literature research started with 104 articles identified, however this number was narrowed down to 19 articles, focusing on the subject in study, that is AI-Driven Cyber Attacks. Also, the search was reduced to the articles published in the last four years.

4.2. CONTENT ANALYSIS

This section describes the content analysis regarding the selected studies, which are the risks and challenges proposed by ai-driven cyber attacks (5.4.1), potential methods to defend against ai-driven cyber attacks (5.4.2), and the future of ai-driven cyber attacks (5.4.3).

4.2.1. Risks and challenges proposed by AI-Driven Cyber Attacks

In this section, we discuss the risks and challenges associated with AI-driven cyber attacks as identified from the articles in our sample. We categorized these challenges into three main areas: 1) data and algorithm challenges, 2) organizational and operational challenges, and 3) ethical and societal challenges (see Table 2).

First, data and algorithm challenges refer to the difficulties in securing and validating the data that AI systems use to detect and respond to cyber threats in IoT environments (Guembe et al., 2022). There are significant risks when AI algorithms are exploited or manipulated to evade detection, thereby launching sophisticated attacks (Guembe et al., 2022). Additionally, issues related to the quality and reliability of data used in AI-based cyber attack detection and mitigation in microgrid systems pose substantial challenges (Beg et al., 2023). Furthermore, dependence on external data sources and potential biases introduced into AI models can impact the effectiveness of cybersecurity measures (Beg et al., 2023).

Second, organizational and operational challenges include resistance and bureaucratic obstacles to implementing AI-driven cybersecurity strategies within Industry 4.0 contexts (Bécue et al., 2021). Integrating AI technologies into existing cybersecurity frameworks and policies within industrial settings also presents significant difficulties (Bécue et al., 2021).

Additionally, governance and regulatory challenges arise as organizations attempt to adapt to the rapid evolution of AI-driven cyber threats, particularly in both public and private sectors (Velasco, 2022). Competing institutional priorities and resource allocation constraints further affect the deployment of effective AI defenses (Velasco, 2022).

Third, ethical and societal challenges involve ethical considerations surrounding the use of AI in cybercrime and the potential societal impacts of malicious AI applications (Blauth et al., 2022). This includes the implications for AI governance and regulatory frameworks needed to mitigate risks associated with AI-driven cyber attacks (Blauth et al., 2022). Societal trust and transparency issues arise from AI-enabled cyber attacks, influencing public perception and policy responses (Yamin et al., 2021). Moreover, the sophisticated nature and global reach of AI-driven attacks pose significant challenges in attributing and responding to these threats (Yamin et al., 2021).

Table 2 - Risks and challenges proposed by AI-Driven Cyber Attacks

Category	Risks and challenges
1) Data and algorithm challenges	Challenges in securing and validating data used by AI systems for detecting and responding to cyber threats in IoT environments. (Guembe et al., 2022) Risks associated with AI algorithms being exploited or manipulated to evade detection and launch sophisticated attacks. (Guembe et al., 2022) Issues related to the quality and reliability of data used in AI-based cyber attack detection and mitigation in microgrid systems. (Beg et al., 2023) Dependence on external data sources and potential biases introduced into AI models, impacting the effectiveness of cybersecurity measures. (Beg et al., 2023)
2) Organizational and operational challenges	Organizational resistance and bureaucratic obstacles to implementing AI-driven cybersecurity strategies in

	Industry 4.0 contexts. (Bécue et al., 2021) Challenges in integrating AI technologies into existing cybersecurity frameworks and policies within industrial settings. (Bécue et al., 2021) Governance and regulatory challenges in adapting to the rapid evolution of AI-driven cyber threats, particularly in public and private sectors. (Velasco, 2022) Competing institutional priorities and resource allocation constraints affecting the deployment of effective AI defenses. (Velasco, 2022)
3) Ethical and societal challenges	Ethical considerations surrounding the use of AI in cybercrime and the potential societal impacts of malicious AI applications. (Blauth et al., 2022) Implications for AI governance and regulatory frameworks to mitigate risks associated with AI-driven cyber attacks. (Blauth et al., 2022) Societal trust and transparency issues arising from AI-enabled cyber attacks, influencing public perception and policy responses. (Yamin et al., 2021) Challenges in attributing and responding to AI-driven attacks due to their sophisticated nature and global reach. (Yamin et al., 2021)

4.2.2. Potential methods to defend against AI-Driven Cyber Attacks

In this section, we outline potential methods to defend against AI-driven cyber attacks, as identified from the literature in our sample. These defense methods are categorized into four main areas: 1) technological and strategic approaches, 2) collaboration and governance strategies, 3) advanced detection and response systems, and 4) AI-based security solutions (see Table 3).

First, technological and strategic approaches involve utilizing AI for proactive threat intelligence and predictive analytics to detect and mitigate emerging cyber threats (Li, 2018). Integration of AI-driven anomaly detection and behavioral analytics into cybersecurity frameworks is also crucial for enhancing detection capabilities (Li, 2018). Furthermore, the development of AI-assisted frameworks for real-time monitoring and response in software-defined mobile networks helps mitigate AI-driven cyber risks (Catania et al., 2018).

Second, collaboration and governance strategies focus on strengthening international collaboration and information sharing mechanisms to combat cross-border AI-driven cyber threats (Kaloudi & Li, 2021). Implementing robust AI governance frameworks and regulatory standards ensures responsible AI use in cybersecurity (Kaloudi & Li, 2021). Promoting public-private partnerships and sectoral collaboration is essential to developing and deploying AI-driven cybersecurity solutions effectively (Werthner et al., 2022). Additionally, enhancing organizational readiness through capacity building and training in AI technologies and cyber defense strategies is critical (Werthner et al., 2022).

Third, advanced detection and response systems leverage AI for automated incident response and remediation in real-time (Bécue et al., 2021). Developing AI-driven intrusion detection systems (IDS) and intrusion prevention systems (IPS) enhances the capability to detect and mitigate cyber threats (Bécue et al., 2021). Machine learning models are also employed for anomaly detection in network traffic and user behavior, bolstering overall security posture (Catania et al., 2018). Moreover, utilizing AI for proactive threat hunting and intelligence gathering strengthens defense mechanisms against evolving threats (Catania et al., 2018).

Fourth, AI-based security solutions involve integrating AI with traditional cybersecurity tools such as firewalls, anti-malware, and encryption to enhance their effectiveness (Li, 2018). The use of reinforcement learning allows these systems to continuously adapt and improve their defenses against sophisticated attacks (Li, 2018). Additionally, the development of AI-based honeypots and deception technologies aids in detecting and studying attacker behavior, providing insights for better defense strategies (Kaloudi & Li, 2021). Furthermore, applying AI in risk assessment and management helps prioritize and mitigate potential threats effectively (Kaloudi & Li, 2021).

Table 3 - Potential methods to defend against AI-Driven Cyber Attacks

Category	Potential defense methods
1) Technological and strategic approaches	Utilization of AI for proactive threat intelligence and predictive analytics to detect and mitigate emerging cyber threats (Li, 2018) Integration of AI-driven anomaly detection and behavioral analytics into cybersecurity frameworks to enhance detection capabilities (Li, 2018) Development of AI-assisted frameworks for real-time monitoring and response in software-defined mobile networks to mitigate AI-driven cyber risks (Catania et al., 2018)
2) Collaboration and governance strategies	Strengthening international collaboration and information sharing mechanisms to combat cross-border AI-driven cyber threats (Kaloudi & Li, 2021) Implementation of robust AI governance frameworks and regulatory standards to ensure responsible AI use in cybersecurity (Kaloudi & Li, 2021) Promoting public-private partnerships and sectoral collaboration to develop and deploy AI-driven cybersecurity solutions effectively (Werthner et al., 2022) Enhancing organizational readiness through capacity building and training in AI technologies and cyber defense strategies (Werthner et al., 2022)
3) Advanced detection and response systems	Leveraging AI for automated incident response and remediation in real-time (Bécue et al., 2021).

	Development of AI-driven intrusion detection systems (IDS) and intrusion prevention systems (IPS) (Bécue et al., 2021). Implementation of machine learning models for anomaly detection in network traffic and user behavior (Catania et al., 2018). Utilization of AI for proactive threat hunting and intelligence gathering (Catania et al., 2018).
4) AI – based security solutions	Integration of AI with traditional cybersecurity tools such as firewalls, anti-malware, and encryption (Li, 2018). Use of reinforcement learning to continuously adapt and improve cybersecurity defenses (Li, 2018) Development of AI-based honeypots and deception technologies to detect and study attacker behavior (Kaloudi & Li, 2021) Application of AI in risk assessment and management to prioritize and mitigate potential threats (Kaloudi & Li, 2021)

4.2.3. Future of AI-Driven Cyber Attacks

In this section, we explore the future trends and implications of AI-driven cyber attacks based on the literature review conducted. The future of AI-driven cyber attacks is categorized into three main areas: 1) emerging trends and predictions, 2) technological evolution and impact, and 3) strategic and policy developments (see Table 4).

First, emerging trends and predictions indicate that AI-driven cyber attacks are expected to evolve towards autonomous threats targeting critical infrastructure and AI-dependent systems (Guembe et al., 2022). Advancements in AI capabilities will likely lead to more sophisticated evasion techniques and rapid adaptation to defensive measures, challenging cybersecurity paradigms (Guembe et al., 2022). There is an anticipated growth in AI-based attack vectors such as AI-driven social engineering and deepfake technologies, which pose significant challenges to traditional cybersecurity defenses (Hassan & Wasim, 2023). This trend raises implications for global cybersecurity policies and regulatory frameworks, necessitating adaptive responses to the escalating threat landscape (Hassan & Wasim, 2023).

Second, technological evolution and impact predictions suggest a convergence of AI with emerging technologies like 5G and IoT, expanding the attack surface for cyber threats (Guembe et al., 2022). The future development of autonomous AI agents capable of launching and defending against cyber attacks without human intervention is also anticipated (Guembe et al., 2022). The rise in AI-driven deepfake attacks is expected to impact information integrity and trust, particularly in digital environments (Hassan & Wasim, 2023). Furthermore, AI's role in enabling cyber-physical attacks targeting smart cities and connected infrastructure is a growing concern (Hassan & Wasim, 2023).

Third, strategic and policy developments include the evolution of global cybersecurity policies to address the specific challenges posed by AI-driven threats (Kaloudi & Li, 2021). Initiatives are underway to develop resilient AI systems capable of withstanding adversarial attacks and ensuring robust cybersecurity defenses (Kaloudi & Li, 2021). Future trends in AI ethics and regulation will shape the development and deployment of secure AI technologies, influencing their adoption across various sectors (Werthner et al., 2022). Moreover, the impact of AI-driven cyber threats on international relations and geopolitical stability is a growing area of concern, requiring strategic foresight and collaboration among nations (Werthner et al., 2022).

In conclusion, the future of AI-driven cyber attacks presents both opportunities and challenges, demanding proactive measures in technology development, policy formulation, and international cooperation to safeguard against emerging threats in cyberspace.

Table 4 - Future of AI-Driven Cyber Attacks

Category	Future of AI-Driven Cyber Attacks
1) Emerging trends and predictions	Future scenarios of AI-driven cyber attacks evolving towards AI-enabled autonomous threats targeting critical infrastructure and AI-dependent systems (Guembe et al., 2022) Potential advancements in AI capabilities leading to more sophisticated evasion techniques and rapid adaptation to defensive measures (Guembe et al., 2022) Anticipated growth in AI-based attack vectors such as AI-driven social engineering and deepfake technologies, challenging traditional cybersecurity paradigms (Hassan & Wasim, 2023) Implications for global cybersecurity policies and regulatory frameworks in response to the escalating threat landscape (Hassan & Wasim, 2023)
2) Technological evolution and impact	Predictions on the convergence of AI with emerging technologies such as 5G and IoT, increasing the attack surface (Guembe et al., 2022) Future development of autonomous AI agents capable of launching and defending against cyber attacks without human intervention (Guembe et al., 2022) Anticipated rise in AI-driven deepfake attacks and their implications for information integrity and trust (Hassan & Wasim, 2023)

	Exploration of AI's role in enabling cyber-physical attacks targeting smart cities and connected infrastructure (Hassan & Wasim, 2023)
3) Strategic and policy developments	Evolution of global cybersecurity policies to address the challenges posed by AI-driven threats (Kaloudi & Li, 2021) Strategic initiatives to develop resilient AI systems capable of withstanding adversarial attacks (Kaloudi & Li, 2021) Future trends in AI ethics and regulation, shaping the development and deployment of secure AI technologies (Werthner et al., 2022) Impact of AI-driven cyber threats on international relations and geopolitical stability (Werthner et al., 2022)

Table 5 - Information collected about the selected studies

Category	Metadata	Description
Descriptive information	Article number (#)	Assigned study number
	Reference in APA style	What is the full reference for this source, including the article's author(s), year of publication, title, and any other relevant source details?
	Keywords	What are the keywords identified in the study?
	Database	Which database was used to find this article?
	Journal / conference	In which journal or conference was the article published?
	Found through (database)	Which database was used to find this article?
Approach-related information	Study objective / Research question	What is the study objective and the main research question?
	Research method(s)	What are the methods used to collect data in the selected studies?
	Qualitative/ quantitative methods / mixed methods	Does the study use qualitative, quantitative or mixed methods?
	Theory mentioned	Does the study mention any theory? If yes, what theory?
	Use of theory	If any theory is mentioned, how is theory used in the study? (e.g. mentioned to explain a certain phenomenon, used as a framework for analysis, tested theory, theory mentioned in the future research section)
Quality-related information	Research approach	Is the design of the study appropriate with respect to the research objectives?

	Quality assessment	Are there any quality concerns? (e.g. limited information about the research methods used)
AI-Driven Cyber Attacks related information	Study's contributions	What are the study contributions stated by the author(s)?
	Risks and challenges proposed by ai-driven cyber attacks	What are the risks and challenges proposed by the use of ai-driven cyber attacks?
	Potential methods to defend against ai-driven cyber attacks	What are the potential methods to defend against ai-driven cyber attacks?

Table 6 - Overview of the objetives from the selected studies

#	Reference	Study objective
1	(Abdullahi et al., 2022)	To categorize, map, and survey AI methods for detecting cybersecurity attacks in IoT environments.
2	(Ansari et al., 2022)	Develop AI-based techniques to detect and prevent phishing attacks.
3	(Babajide Tolulope Familoni, 2024)	To explore and elucidate the intricate theoretical and practical dimensions of addressing cybersecurity challenges in the age of artificial intelligence (AI), with a focus on understanding the complex interplay between AI algorithms, human behavior, and adversarial tactics.
4	(Bécue et al., 2021)	To discuss opportunities and threats of using AI in the manufacturing sector, focusing on security principles and detection techniques for operational technology.
5	(Beg et al., 2023)	To present AI-based techniques for cyber-attack detection and mitigation in microgrids and explore future research directions.
6	(Blauth et al., 2022)	To provide a comprehensive understanding of the malicious use and abuse of AI, including the corresponding risks.
7	(Catania et al., 2018)	To demonstrate how AI can support network managers in detecting and

		handling cyber-attacks in mobile networks using the Software Defined Networking (SDN) paradigm.
8	(Comiter, 2019)	To explore the systematic vulnerabilities of AI systems to "artificial intelligence attacks" and their potential impact on critical components of society.
9	(Guembe et al., 2022)	Identify the current and emerging AI-driven techniques used by malicious attackers to carry out cybercrime.
10	(Hassan & Wasim, 2023)	To examine the intersection between AI and digital security, exploring AI's role in enhancing cybersecurity and addressing potential threats posed by hostile AI.
11	(Kaloudi & Li, 2021)	To explore and classify AI-based cyber attacks, providing a framework for understanding and detecting these threats to predict future risks.
12	(Li, 2018)	To review the intersection of AI and cybersecurity, summarizing AI methods for combating cyberattacks and analyzing the counterattacks and defenses for AI systems.
13	(Velasco, 2022)	To evaluate international instruments and propose policy responses for countering AI-enabled cybercrimes.
14	(Werthner et al., 2022)	To challenge the assumption of top-down, intelligent design in

		technology development and propose a Darwinian evolutionary perspective where humans act as agents of mutation.
15	(Yamin et al., 2021)	To investigate AI-based cyberattacks, identify mitigation strategies, and anticipate future scenarios to enhance defense against such attacks.

5. DISCUSSION

The landscape of cybersecurity is rapidly evolving with the advent of AI-driven cyber attacks. Existing literature often focuses on specific aspects of these attacks, such as emerging threats, specific vulnerabilities, or particular mitigation strategies. However, there is a need for a comprehensive study that encapsulates the entire spectrum of AI-driven cyber attacks. This thesis aims to fill that gap by providing a holistic overview of AI-driven cyber attacks, examining the entire lifecycle from threat emergence to mitigation strategies and ethical implications. This study differentiates itself by integrating insights from multiple dimensions of AI-driven cyber attacks, thereby offering a more complete picture. Unlike previous studies that may concentrate on isolated facets, this research encompasses a broad range of topics, including the mechanisms of AI-driven attacks, the diverse vulnerabilities they exploit, the various defense mechanisms available, and the overarching ethical and legal considerations. By doing so, this thesis aims to provide a more comprehensive understanding of the AI-driven cyber threat landscape and to guide future research and policy development in this critical area.

Based on the findings of this thesis, 15 practical recommendations can be made to enhance the cybersecurity landscape in the face of AI-driven threats:

1) Enhancing Cybersecurity Training

Regular training programs for cybersecurity professionals should include specialized modules on AI-driven cyber attacks and defense mechanisms. These programs should cover the latest developments in AI technology and its application in cyber threats, as well as hands-on exercises to recognize and mitigate such threats. Awareness campaigns targeting the general public can also play a crucial role in reducing the risk of social engineering attacks by educating users on safe online practices and the signs of AI-driven phishing attempts.

2) Developing Robust Defense Mechanisms

Investment in research and development of advanced defense mechanisms is essential. Collaboration between academia, industry, and government can foster innovation in cybersecurity tools and techniques. For example, developing AI-based anomaly detection systems that can identify and respond to threats in real time can significantly enhance defense capabilities. Additionally, creating a shared repository of threat intelligence can help organizations stay informed about the latest attack vectors and defense strategies.

3) Strengthening Legal and Ethical Frameworks

Policymakers should work towards establishing comprehensive legal and ethical frameworks to regulate the use of AI in cybersecurity. These frameworks should balance the need for robust security measures with the protection of individual privacy and civil liberties. International cooperation is crucial to address the global nature of AI-driven cyber threats,

and establishing common standards and protocols can facilitate more effective cross-border collaboration.

4) Encouraging Interdisciplinary Research

Interdisciplinary research involving experts from computer science, law, ethics, and social sciences can provide holistic solutions to the challenges posed by AI-driven cyber attacks . Such collaboration can lead to the development of AI defense systems that are not only technically effective but also aligned with ethical principles and legal requirements. For example, a multidisciplinary team could design an AI-based monitoring system that ensures user privacy while effectively detecting malicious activities.

5) Fostering Public-Private Partnerships

Public-private partnerships can play a significant role in enhancing cybersecurity. Governments can provide funding and regulatory support, while private companies can contribute technological expertise and resources . Such partnerships can lead to the development of innovative solutions, such as shared cybersecurity infrastructure and collaborative threat intelligence platforms .

6) Promoting Cyber Hygiene Practices

Encouraging organizations and individuals to adopt good cyber hygiene practices can significantly reduce the risk of AI-driven cyber attacks . This includes regular software updates, strong password policies, and the use of multi-factor authentication. Educating users about the importance of these practices and providing easy-to-follow guidelines can enhance overall cybersecurity resilience .

7) Implementing AI Ethics Guidelines

Developing and implementing AI ethics guidelines can help ensure that AI technologies are used responsibly in cybersecurity . These guidelines should address issues such as bias in AI algorithms, transparency in AI decision-making, and accountability for AI-driven actions. Adhering to ethical standards can enhance public trust in AI-based cybersecurity solutions .

8) Investing in Cybersecurity Research and Development

Governments and organizations should invest in cybersecurity research and development to stay ahead of AI-driven cyber threats. Funding innovative projects and supporting cutting-edge research can lead to the discovery of new defense mechanisms and the improvement of existing ones. This investment can also foster the development of a skilled cybersecurity workforce equipped to handle advanced AI-driven threats.

9) Enhancing Incident Response Capabilities

Improving incident response capabilities can help organizations quickly and effectively respond to AI-driven cyber attacks . This includes developing comprehensive incident response plans, conducting regular drills and simulations, and establishing clear communication channels for reporting and managing incidents. Rapid response can mitigate the impact of attacks and facilitate faster recovery .

10) Strengthening Threat Intelligence Sharing

Encouraging organizations to share threat intelligence can enhance collective cybersecurity defenses. Creating secure platforms for sharing information about AI-driven cyber threats, attack patterns, and defense strategies can help organizations stay informed and better prepared to counter emerging threats. Collaboration among industry peers and with government agencies can improve overall situational awareness.

11) Developing Adaptive Security Architectures

Adopting adaptive security architectures that can dynamically adjust to changing threat landscapes is crucial in countering AI-driven cyber attacks . These architectures should leverage AI and machine learning to continuously monitor and analyze network traffic, detect anomalies, and respond to threats in real time. Implementing adaptive security measures can enhance the resilience of cybersecurity systems .

12) Conducting Regular Security Audits

Regular security audits can help organizations identify and address vulnerabilities before they are exploited by AI-driven cyber-attacks. These audits should include thorough assessments of network infrastructure, software applications, and security policies. Implementing recommendations from security audits can strengthen an organization’s defense posture.

13) Emphasizing Proactive Threat Hunting

Proactive threat hunting involves actively seeking out potential threats within an organization’s network before they can cause harm. Utilizing AI-driven tools for threat hunting can enhance the ability to detect and neutralize advanced threats. Training cybersecurity teams in threat hunting techniques and providing them with the necessary tools and resources can improve overall security.

14) Supporting Cybersecurity Education

Promoting cybersecurity education at all levels can help build a knowledgeable and skilled workforce capable of addressing AI-driven cyber threats. This includes integrating cybersecurity topics into school curricula, offering specialized courses and certifications, and providing opportunities for hands-on learning and practical experience.

15) Advocating for Responsible AI Development

Advocating for responsible AI development practices can help mitigate the risks associated with AI-driven cyber-attacks. Encouraging developers to follow best practices in AI design, testing, and deployment can reduce vulnerabilities and ensure that AI technologies are used ethically and securely. Promoting awareness of AI risks and responsible development among the tech community can contribute to a safer digital environment.

6. CONCLUSIONS AND FUTURE WORKS

This research has highlighted the significant risks and challenges posed by AI-driven cyber attacks. Through comprehensive analysis, it has been demonstrated that the integration of AI into cyber attack methodologies has led to more sophisticated, adaptive, and autonomous threats. The findings emphasize the urgent need for advanced cybersecurity measures capable of countering these evolving threats. Key conclusions from this study include the increased sophistication and autonomy of AI-driven attacks, which leverage advanced algorithms to bypass traditional security measures, making them more effective and harder to detect and mitigate. Moreover, the use of AI has introduced new attack vectors such as AI-driven social engineering and deepfake technologies, posing significant challenges to existing cybersecurity frameworks. The study underscores the necessity for developing advanced defensive strategies that incorporate AI for detecting and mitigating threats, including the use of machine learning models that can predict and respond to attack patterns in real-time. Additionally, the rapid evolution of AI in cyber attacks necessitates a robust regulatory framework to address ethical concerns and ensure the responsible use of AI technologies.

BIBLIOGRAPHICAL REFERENCES

Abdullahi, M., Baashar, Y., Alhussian, H., Alwadain, A., Aziz, N., Capretz, L. F., & Abdulkadir, S. J. (2022). Detecting Cybersecurity Attacks in Internet of Things Using Artificial Intelligence Methods: A Systematic Literature Review. *Electronics*, *11*(2), 198. https://doi.org/10.3390/electronics11020198

Anderson, R., Barton, C., Böhme, R., Clayton, R., Van Eeten, M. J. G., Levi, M., Moore, T., & Savage, S. (2013). Measuring the Cost of Cybercrime. In R. Böhme (Ed.), *The Economics of Information Security and Privacy* (pp. 265–300). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39498-0_12

Ansari, M. F., Sharma, P. K., & Dash, B. (2022). Prevention of Phishing Attacks Using AI-Based Cybersecurity Awareness Training. *International Journal of Smart Sensor and Adhoc Network.*, 61–72. https://doi.org/10.47893/IJSSAN.2022.1221

Babajide Tolulope Familoni. (2024). CYBERSECURITY CHALLENGES IN THE AGE OF AI: THEORETICAL APPROACHES AND PRACTICAL SOLUTIONS. *Computer Science & IT Research Journal*, *5*(3), 703–724. https://doi.org/10.51594/csitrj.v5i3.930

Bano, M., & Zowghi, D. (2015). A systematic review on the relationship between user involvement and system success. *Information and Software Technology*, *58*, 148–169. https://doi.org/10.1016/j.infsof.2014.06.011

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, *41*(3), 16:1-16:52. https://doi.org/10.1145/1541880.1541883

Bécue, A., Praça, I., & Gama, J. (2021). Artificial intelligence, cyber-threats and Industry 4.0: Challenges and opportunities. *Artificial Intelligence Review*, *54*(5), 3849–3886. https://doi.org/10.1007/s10462-020-09942-2

Beg, O., Khan, A., Rehman, W., & Hassan, A. (2023). A Review of AI-Based Cyber-Attack Detection and Mitigation in Microgrids. *Energies*, *16*(22), 7644. https://doi.org/10.3390/en16227644

Bilge, L., & Dumitraş, T. (2012). Before we knew it: An empirical study of zero-day attacks in the real world. *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, 833–844. https://doi.org/10.1145/2382196.2382284

Blauth, T. F., Gstrein, O. J., & Zwitter, A. (2022). Artificial Intelligence Crime: An Overview of Malicious Use and Abuse of AI. *IEEE Access*, *10*, 77110–77122. https://doi.org/10.1109/ACCESS.2022.3191790

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G., Steinhardt, J., Flynn, C., hÉigeartaigh, S., Beard, S., Belfield, H., Farquhar, S., & Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*.

Catania, G., Ganga, L., Milardo, S., Morabito, G., & Mursia, A. (2018). *An AI-Assisted Cyber Attack Detection Framework for Software Defined Mobile Networks*.

Citron, D., & Chesney, R. (2019). Deepfakes and the New Disinformation War. *Foreign Affairs*. https://scholarship.law.bu.edu/shorter_works/76

Comiter, M. (2019). *Attacking Artificial Intelligence*.

Egele, M., Scholte, T., Kirda, E., & Kruegel, C. (2008). A survey on automated dynamic malware-analysis techniques and tools. *ACM Comput. Surv.*, *44*(2), 6:1-6:42. https://doi.org/10.1145/2089125.2089126

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, Inc.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). The Emerging Threat of Ai-driven Cyber Attacks: A Review. *Applied Artificial Intelligence*, *36*(1), 2037254. https://doi.org/10.1080/08839514.2022.2037254

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, *24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36

Hassan, S. M., & Wasim, D. J. (2023). *STUDY OF ARTIFICIAL INTELLIGENCE IN CYBER SECURITY AND THE EMERGING THREAT OF AI-DRIVEN CYBER ATTACKS AND CHALLENGES*.

Hong, J. (2012). The State of Phishing Attacks. *Commun. ACM*, *55*, 74–81. https://doi.org/10.1145/2063176.2063197

Kaloudi, N., & Li, J. (2021). The AI-Based Cyber Threat Landscape: A Survey. *ACM Computing Surveys*, *53*(1), 1–34. https://doi.org/10.1145/3372823

Kitchenham, B. (2004). *Procedures for Performing Systematic Reviews*.

Kshetri, N. (2018). The Economics of Cyber-Insurance. *IT Professional*, *20*(6), 9–14. https://doi.org/10.1109/MITP.2018.2874210

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, J. (2018). Cyber security meets artificial intelligence: A survey. *Frontiers of Information Technology & Electronic Engineering*, *19*(12), 1462–1474. https://doi.org/10.1631/FITEE.1800573

NIST. (2018). *Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1* (NIST CSWP 04162018; p. NIST CSWP 04162018). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.CSWP.04162018

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. *2016 IEEE Symposium on Security and Privacy (SP)*, 582–597. https://doi.org/10.1109/SP.2016.41

Russell, S. J., Norvig, P., Chang, M., Devlin, J., Dragan, A., Forsyth, D., Goodfellow, I., Malik, J., Mansinghka, V., Pearl, J., & Wooldridge, M. J. (2022). *Artificial intelligence: A modern approach* (Fourth edition, global edition). Pearson.

SANS Institute. (2020). https://www.sans.org/security-awareness-training/

Stallings, W. (2017). *Cryptography and network security: Principles and practice* (Seventh edition). Pearson.

Tankard, C. (2011). Advanced Persistent threats and how to monitor and deter them. *Network Security*, *2011*(8), 16–19. https://doi.org/10.1016/S1353-4858(11)70086-1

Velasco, C. (2022). Cybercrime and Artificial Intelligence. An overview of the work of international organizations on criminal justice and the international applicable instruments. *ERA Forum*, *23*(1), 109–126. https://doi.org/10.1007/s12027-022-00702-z

Werthner, H., Prem, E., Lee, E. A., & Ghezzi, C. (Eds.). (2022). *Perspectives on Digital Humanism*. Springer International Publishing. https://doi.org/10.1007/978-3-030-86144-5

Yamin, M. M., Ullah, M., Ullah, H., & Katt, B. (2021). Weaponized AI for cyber attacks. *Journal of Information Security and Applications*, *57*, 102722. https://doi.org/10.1016/j.jisa.2020.102722

Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, *38*(3), 101577. https://doi.org/10.1016/j.giq.2021.101577

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa