AI-Driven Cyberattack Risks - A Foresight Study

Leon Helgeland
University of Agder, Norway,
Email leon.he@outlook.com

Alin Napirca
University of Agder, Norway,
Email n.alin11@yahoo.com

Jaziar Radianti *
University of Agder, Norway,
Email jaziar.radianti@uia.no

ABSTRACT

The digitalization has integrated artificial intelligence (AI) across multiple domains, including cybersecurity, creating both opportunities and risks. While AI enhances organizational defense through automated threat detection and incident response, it also enables scalable and autonomous cyberattacks by threat actors. This study examines future risks posed by AI-driven cyberattacks using foresight methodologies that combine horizon scanning and scenario development to identify current and emerging offensive AI capabilities. Previous studies indicates that present AI-enabled threats primarily involve social engineering and adversarial attacks, while future risks may include quantum-powered AI and fully autonomous systems. The study highlights a widening gap between offensive and defensive AI, driven by incentives, technical skill, and expertise. Three cyberattack scenarios involving AI tools were developed and validated through interviews with experts to assess plausibility and future relevance. Finally, the study proposes precautionary actions to strengthen organizational resilience, emphasizing anticipatory governance and cross-industry collaboration to improve threat intelligence accuracy

Keywords

AI-Driven Cyberattacks, Foresight, Risks, Cyber Defense

INTRODUCTION

Nowadays, there is a growing overlap between emergency management and cybersecurity, driven by digital transformation and increasing reliance on technology in critical services (Radianti 2024; Radianti 2025). The article highlights this connection, noting that the distinction between the two domains is becoming less clear as technology becomes integral to emergency response, for example through the adoption of tools for data collection, exchange, and storage (Radianti 2024). The digitalization of infrastructure supporting critical services has also transformed the threat landscape, as cybersecurity incidents can now trigger emergency management responses, such as cyberattacks on transport control systems causing citywide traffic disruption. Overall, the article emphasizes the importance of professionals in both cybersecurity and emergency management understanding emerging AI-driven cyberattack risks.

As artificial intelligence becomes more accessible to public, so do the malicious capabilities of individuals, organizations, and states. AI can enhance the security of digital infrastructure but it also enables attackers to launch cyberattacks easier. AI depicts a double-edged sword: it can strengthen organizational cyber-defenses, at the same time while simultaneously it ”upgrades” the malicious capabilities of cyber-criminals. Today´s AI-enabled attacks are more scalable, adaptive, and autonomous, thus existing security frameworks are lagging behind. The European Union Agency for Cybersecurity (ENISA)’s Threat Landscape Report (2024) refers to AI-driven cyberattacks as a growing concern, as the report reveals cyberattacks such as phishing, malware development, and vulnerability exploitation have employed AI (ENISA 2024b). Growing trends of AI-driven cyberattacks has introduced a level of

*corresponding author

uncertainty where traditional risk assessment frameworks cannot fully anticipate these emerging trends. Foresight methodologies have advantages as it is open for expert opinions and other trends identified in written sources to anticipate plausible futures, identifying emerging risks, and supporting long-term planning in dynamic domains such as cybersecurity (UNDP 2018).

This concern is shared by many security organizations such as ENISA, that also publishes Cybersecurity Threats Foresight 2030 report covering current and future risks that involves offensive AI usage (2024). The report outlines likely future scenarios steered by malicious AI use, underpinning the need for cybersecurity professionals to pre-empt both known threats and unknown threats yet to emerge (ENISA 2024a). There is a need to understand the capabilities of current AI. However, at the speed that AI is currently evolving it also becomes important to understand the future risks, consequences and challenges AI-driven operations bring. Moreover, in many organizations and businesses, they lack of security professionals and staff proficient with AI(Poremba 2025), which makes the this article topic even more important to comprehend current threats and formulate realistic scenario-based events to proactively anticipate future AI-driven attacks.

Research on how cyber defense capabilities can adapt to emerging AI-driven threats remains limited. Conventional threat modeling approaches (UK National Cyber Security Centre 2023) often struggle with uncertainty; therefore, this study adopts a foresight approach to explore future risks posed by AI-enabled cyberattacks. The methodology combines horizon scanning, scenario development, and expert interviews. This study aims to provide researchers and cybersecurity professionals and policymakers with knowledge on actionable guidance for preparing against future AI capabilities. The foresight framework is applied (UNDP 2018) to generate insights into future threat capabilities. Guided by the research question—*How can foresight analysis of future AI capabilities inform proactive cybersecurity strategies to help professionals prepare for emerging AI threats?*—the study examines current offensive and defensive AI trends and develops future scenarios through horizon scanning. The scope of this study focuses solely on AI-based threats, and excludes non-AI attack vectors. While previous studies on both offensive and defensive AI are reviewed, the primary emphasis is on building overview on current and future offensive AI capabilities.

LITERATURE REVIEW

This review uses a systematic literature review approach, not only to summarize previous studies on offensive and defensive uses of AI, but also to support horizon scanning for developing scenarios for the foresight activity. To identify high-quality literature, two databases with advanced filtering and strong publishing standards were selected: **AIS eLibrary** (Senior Scholars' List of Premier Journals, emphasizing rigorous peer review in Information Systems and Information Security) (AIS n.d.) and **Scopus** (broad, widely used coverage). Searches used the Boolean string

> ("artificial intelligence" OR "AI") AND ("cyberattack" OR "cyber attack" OR "malware" OR "ransomware") AND ("cyber defense" OR "cybersecurity" OR "cyber security" OR "information security")

and were limited to 2020–2025 to capture current and emerging trends relevant for foresight analysis. The query returned 305 records, which were screened using PRISMA-aligned inclusion and exclusion criteria (Page et al. 2021). Records were included only if accessible (Open Access or through university services), written in English, substantively addressed AI in cyber defense/attack or cybersecurity development, and met quality checks via Kanalregisteret where available (with additional validation for venues not listed) (Norwegian Directorate for Higher Education and Skills 2024). After title screening, 127 articles remained for abstract screening, and 47 remained for full-text eligibility assessment. In total, 18 studies were included in the review. This section is organized based on themes appeared in the previous studies.

Studies on Offensive AI

The reviewed literature shows that AI-enabled cyberattacks range from AI-generated phishing and malicious payloads, adversarial attacks targeting AI models, and fully autonomous malware capable of target selection, lateral movement, evasion, and data exfiltration.

Generative AI: Generative AI has transformed social engineering by enabling synthetic media such as deepfakes and highly convincing phishing campaigns. Advances in GANs (Generative Adversarial Networks) allow attackers to create impersonation content that is nearly indistinguishable from real media (Kazimierczak et al. 2024). Schmitt et al. (Schmitt and Flechais 2023) propose an AI-enabled social engineering framework based on realistic content generation, targeted personalization, and automated attack infrastructure, illustrating AI's role across the entire

attack chain. Beyond social engineering, LLMs ((Large Language Models) are also used to generate executable code and malware for various attack vectors (Iturbe et al. 2024), significantly lowering the barrier for less skilled attackers to launch high-impact attacks.

Adversarial Attacks: The attacks exploit vulnerabilities in machine learning models to compromise confidentiality, integrity, or availability, either during training or at inference time (Mirsky et al. 2023). Techniques such as data poisoning attacks, equation solving attacks, path finding attacks, and AI model inversion attacks can manipulate learning processes or data used by AI systems, leading to misclassifications and faulty outcomes (Malatji 2023). A key factor behind successful attacks is the lack of adequate threat modeling against adversarial Machine Learning, where attackers explicitly target model defenses (Arp et al. 2021). These attacks predominantly target black-box models (Kazimierczak et al. 2024), whose limited explainability further weakens defensive capabilities in cybersecurity contexts.

Autonomous AI Attacks: The attacks employ automated techniques to discover and exploit system vulnerabilities, significantly scaling cyberattacks by increasing their frequency, reach, and number of participating actors (Kamoun et al. 2020). LLMs can enhance brute-force and credential-stuffing attacks by leveraging prior knowledge, while AI systems can intelligently discover insecure web paths using semantic embeddings to infer accessible or sensitive URLs (Castagnaro et al. 2024). Recent deep-learning algorithms enable AI-driven cyber capabilities to adaptively target specific individuals, organizations, or even governments (Nobles 2023; Kazimierczak et al. 2024). The resulting malware can be selectively activated only on identified targets, minimizing collateral damage and reducing detection likelihood. These smart malware systems can autonomously deploy, monitor target environments, and self-initiate attacks when vulnerabilities are detected to maximize impact (Kazimierczak et al. 2024). Additionally, they can evade security controls by mimicking benign behavior and covertly communicate with decentralized command-and-control infrastructures, enabling large-scale botnet operations while complicating attribution and mitigation efforts (Kamoun et al. 2020; Kazimierczak et al. 2024).

Offensive Intelligence Gathering: AI is increasingly used for reconnaissance and target discrimination to improve the precision and success of phishing and malware attacks (Kazimierczak et al. 2024). Through OSINT (Open Source Intelligence), AI—particularly generative models—enables the creation of highly tailored spear-phishing campaigns, including text and deepfake media, to enhance credibility. AI-driven reconnaissance can also uncover exposed or misplaced information online, facilitating credential theft and data exfiltration (Mirsky et al. 2023). While intelligence gathering alone may be benign, its integration with other offensive AI techniques significantly amplifies its destructive potential.

Studies on Defensive AI

Defensive uses of AI were also prominent in prior studies, spanning both proactive and reactive strategies. Most research focused on employing AI for anomaly detection, while also highlighting its role in risk assessment and management, incident response, and post-incident analysis.

AI Detection tools: Organizations employ AI to improve cyber defense capabilities such as threat analysis, automated defense mechanisms, and heightened security postures (Mirsky et al. 2023). This is related to the notion of using ”AI to fight AI” which is a necessary approach to encounter the impact of advancements and sophistication of AI-driven attacks on cyber environments (Guembe et al. 2022). AI-powered countermeasures are mainly autonomous security systems used to detect attacks and respond to breaches. For example, AI-based techniques to improve malware detection capabilities and analyze patterns such as code behavior and API calls to identify malicious software (Sen et al. 2022). Humans are often prone to errors, when handling large volume of data such as frequent security alerts, which can be overwhelming for security teams. AI systems can automatically detect and effectively respond to, for both known and unknown malicious traces (Sen et al. 2022). However, due to the black-box nature of many AI models, AI-powered tools are best combined with traditional cybersecurity solutions to preserve explainability (Malatji 2023). Achieving this human–AI symbiosis requires integrating AI with human oversight, enabling collaborative decision-making that enhances effectiveness and trust (Nobles 2023).

Risk Assessment and Risk Management: The offensive threat of AI can also be utilized to strengthen defense capabilities by using AI to automatically generate attack trees to understand all possibilities that can lead to system failure to help security teams identify vulnerabilities in their systems (Kazimierczak et al. 2024). Security teams can assess the system’s security levels and simulate attacks in a controlled environment (sandbox) to complement Breach and Attack Simulations (BASs) platforms and further strengthen defense capabilities (Iturbe et al. 2024). The test results can be analyzed using AI technologies to automate vulnerability identification and assessment, calculate risk scores in accordance with established frameworks such as ISO 27005 (*Information security, cybersecurity*

and privacy protection — Guidance on managing information security risks 2022), and support decision making through the analysis of log data and threat intelligence (Kaur et al. 2023).

Incident Response: Automated response mechanisms increasingly rely on AI systems to streamline incident response by creating dynamic case management records, executing predefined response protocols, and continuously updating contingency plans based on historical security incident data. (Kaur et al. 2023).

Studies on Future AI Development

Prior studies also discuss the development of future AI tools, such as development frameworks, ethical considerations, and the need for human-AI cooperation. Moreover, some scholars scrutinize themes such as strengthening (cyber) security capabilities using blockchain and Quantum powered AI-systems.

Safe AI: As AI development advances, increasing emphasis is placed on frameworks and methodologies to create safe and effective AI. Recent research indicates a shift from model-centric approaches—focused on algorithmic and computational improvements—to data-centric approaches that prioritize data quality. Kumar et al. (2024) shows that data-centric models often yield greater performances, pointing the importance of robust data collection, data sanitizing, and integration (Kumar et al. 2024). These challenges become more pronounced as development moves beyond *narrow AI* applications toward Artificial General Intelligence (AGI), which aims to transfer knowledge across domains similarly to humans (Nobles 2023, pp. 351, 352). This further reinforce the need for high-quality and well-labeled datasets.

Ethics in AI Development: Ethical considerations surrounding the development and use of AI—particularly in cybersecurity—present significant challenges. The dual-use nature of AI, enabling both offensive and defensive applications, raises major concerns: offensive AI can be exploited by malicious actors, while defensive AI often suffers from limited transparency and explainability due to its black-box decision-making processes (Kaur et al. 2023). As the AI systems become more prevalent, there is a strong need for ethical and regulatory frameworks. Concerns such as the impact of AI on jobs and society, and its accountability increase the need for appropriate regulations and ethical design to ensure responsible AI behavior in cybersecurity (Nobles 2024). Moreover, the use of adversarial attacks to disrupt the AI training process is also prevalent, and policy makers talk about working on modifying existing cybersecurity laws such as $USA's Computer Fraud and Abuse Act$ to create disincentives for such research (Sen et al. 2022).

Human-AI Symbiosis: The complex relationship between humans and AI systems is constantly evolving, and as the complementary strengths of humans and machines can work together it also brings forth some challenges and risks. Effective human-AI symbiosis requires that AI tools cooperate alongside human expertise, as the efficiency of AI tools is undeniable, however, in combination with human judgment a thorough collaborative verification process will yield the best results (McIntosh et al. 2023). Malicious actors already make use of AI-enabled bots to enhance cyberattack capabilities, so defending security teams will be forced to defend with bots as well (Mirsky et al. 2023). Guembe et al. warns however, that at the speed cyber threats evolve, existing cyber defense infrastructures will become inadequate to address these AI-driven attacks and suggests that the human-AI relationship in cybersecurity will need to evolve quickly (Guembe et al. 2022).

Blockchain security: Blockchain is proposed as a data verification strategy for Internet of Things (IoT) devices (Kazimierczak et al. 2024). Within the context of command-and-control defense mechanisms, (2018), a three-level blockchain architecture that integrates AI-based defense mechanisms during the installation phase is proposed to strengthen overall security (Kazimierczak et al. 2024).

(AI) Quantum Computing: Several studies address security challenges that might arise as post-quantum technologies fully emerges. McIntosh et al. and Schmitt et al. argue that stronger authentication will be required in post-quantum era, as AI can threaten the security protocols and encryption methods.(McIntosh et al. 2023; Schmitt and Flechais 2023). Emerging technologies and their implications for social engineering attacks, for creating deepfakes and strong brute-force attacks encourage the need for the development of quantum-resistant encryption for defensive purposes, and potentially new forms of attacks that current systems are not designed to handle (Schmitt and Flechais 2023).

Gaps in the Literature

The literature consistently identifies a widening gaps between offensive and defensive AI capabilities in cybersecurity. Offensive AI has reshaped the threat landscape into an increasingly asymmetric contest, where human-driven and reactive defenses are becoming insufficient. Several scholars argue that defenders are gradually losing the "AI war" (Nobles 2023). This gaps is evident across multiple dimensions. From a technical perspective, attackers exploit

AI to automate sophisticated evasion, reconnaissance, and lateral movement techniques that bypass traditional security controls. Defenders face a structural disadvantage, as protecting against unknown and rapidly evolving threats is complex, costly, and difficult to scale. In contrast, attackers benefit from greater unpredictability, access to computing power, and financial resources, which further amplifies the asymmetry (Guembe et al. 2022; Mirsky et al. 2023).

The gap is worsened by a shortage of expertise. Cybersecurity organizations increasingly lack professionals with sufficient AI and machine learning knowledge to counter offensive AI techniques. This skills deficit limits the development of proactive defenses and slows the design of effective countermeasures (Malatji 2023; Sen et al. 2022).

Attackers also retain the initiative. They determine when, where, and how attacks occur, while defenders remain largely reactive. Survey evidence shows that many security professionals perceive AI as more advantageous for attackers, as it enhances multiple stages of the attack lifecycle—particularly reconnaissance, weaponization, and delivery—while enabling rapid and parallel evolution of attack techniques (Mirsky et al. 2023).

Beyond capability gaps, the literature highlights important research shortcomings. Kaur et al.(2023) note the absence of proprietary or shared platforms for exchanging emerging threat intelligence. This issue is reinforced by limited interdisciplinary research, as most studies on AI-driven cybercrime focus narrowly on technical mechanisms while neglecting social, economic, and political factors. Combined with the shortage of AI-proficient security professionals, these gaps hinder the development of robust defensive AI. Against this background, this foresight study examines current and emerging AI-enabled threats to inform and strengthen future cybersecurity postures.

METHODOLOGY

Foresight

This study adopts a foresight methodology (UNDP 2018), recognizing that the future—especially in cybersecurity—is not predetermined and cannot be reliably addressed through traditional prediction methods. Foresight differs from prediction in its approach to the future. Prediction plans for a single, most likely outcome based on existing knowledge and linear progression, whereas foresight uses scenario planning to explore multiple possible futures. By accounting for uncertainties and unknowns, foresight enables preparation across a range of plausible outcomes rather than relying on one forecast. Figure 1 visualizes the differences:

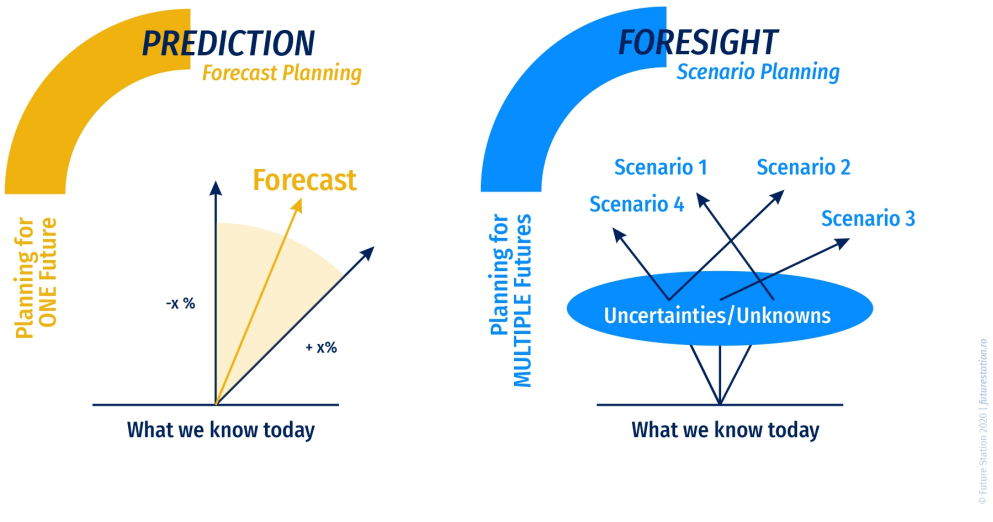

Figure 1. Foresight vs. Prediction visualization

Foresight combines qualitative and quantitative analysis to support long-term planning. See Figure 2 on the variation of foresight methodology. Accordingly, this study employed horizon scanning as a qualitative approach alongside literature review, weak signals, and wild cards to develop future scenarios (Schlagwein et al. 2025; Popper 2008).

The foresight in this study has the following objectives: 1) to create AI cybersecurity threat scenarios based on the horizon scan by identifying weak signals, wild cards, and uncertainties; 2) to identify the plausibility of the

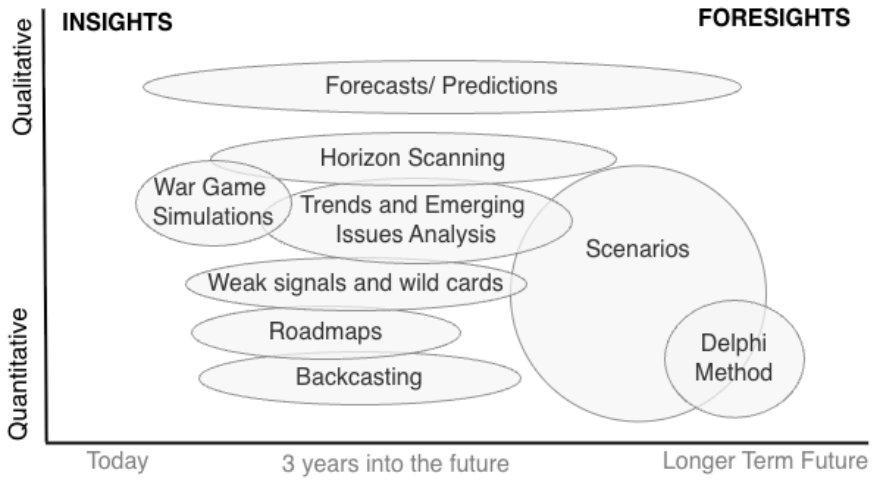

Figure 2. Variation of foresight methodologies (Jacobsen and Hirvensalo 2019).

scenarios based on literature and feedback from experts within the fields; 3) to translate the foresight outputs into strategic implications and policy-relevant recommendations to enhance the cyber defense against the upcoming AI threats. See the procedure application in this study in next section on "Foresight Structure of this Study and Scenario Development)".

Data Collection

There are some alternative qualitative data collection for foresight studies (Popper 2008). However, this study employed semi-structured interviews with experts in combination with scenario scenario plausibility discussion, which allowed us extracting detailed knowledge, opinions, and expectations about future developments from the experts (Giaoutzi and Sapio 2013). The interview approach was semi-structured, combining both theme-specific and open-ended questioning. This structure was chosen in order to maintain some coherence across the different participants as they had different views and expertise, and allowed for freedom of elaboration in areas they deemed significant. This format is also consistent with foresight methods, which prioritize flexibility and adaptivity during knowledge co-creation (Giaoutzi and Sapio 2013, pp. 4–5, 133–134).The scenarios were presented via slide presentation to facilitate visual aid for the conversation and to help participants anchor their attention to the details while engaging in conversation, as suggested by Giaoutzi and Sapio (2013, pp. 177–180). The main eligibility criteria for the participants was expertise within the field of AI or cybersecurity, preferably both. Scenario discussions with experts occurred at the end of each interview. The main goal of introducing the scenario context was to reflect on the plausibility of each scenario event, arguing for why or why not the event is possible. On the details of how the expert interviews included scenario discussions, please see Foresight Structure and Scenario Development Section.

Interviewee Profiles

- *AI-expert-1*: The Expert-1 is dealing with hardware AI accelerators and IoT AI deployments, but is not specifically an AI modeler. The interviewee considers cybersecurity is attractive and compelling theme but considered himself as having knowledge on it in general level

- *AI-expert-2*: Expert-2 is a professor in big data and emerging technology, and has a strong background in AI, has a PhD in AI and has knowledge on semantic technologies. Expert-2 recently works more with AI, and has a lot of experience in AI, including data stream management in AI, social cybersecurity research and also expert on transport and mobility domain. Developing AI tools for data stream management and social cybersecurity.

- *AI-expert-3*: Expert-3 is an AI researcher that focuses on two parts, i.e., AI applications and AI research itself. For the AI application, the interviewee uses deep neural networks and transformer based strategies for some computer vision tasks. But that is only application oriented research, tuning, and parameter changing and cascading some structures of transformers based strategies, but not completely modifying the transformers.

The expert Edge-3 works on AI research mainly, and also have interests on low energy-AI and interpretable machine learning.

- *AI-expert-4*: Expert 4 conducts research on AI for identifying incidents through social media. The expert 4 uses LLMs and machine learning for analysis and predictions, not only limited to cybersecurity, but also prediction of crises through available data, mostly social media data, or bigger events like elections. The expert-4 investigates the framing of messages or videos that were identified as propaganda or misinformation is also one of the things the interviewee uses AI for. Expert-4 has conducted testing some xAI, and classification of images with machine learning.

Data Analysis

The data analysis process consisted mostly of categorization and conceptualization of the information gathered through the interview process. The categorization was done through the identified themes in during the literature review process, and conceptualization through scenario discussion to invoke critical thinking and arguments for and against the plausibility of the created scenarios. Some adjustments sometimes required when conducting data analysis from the interview such as data standardization (Zakaria and Whitfield 2025) of the plausibility of the scenario plausibility. The interviewees each had their own opinion and metric on how to grade the plausibility of the scenarios, which made their answers not standardized. The interviews has brought forward a lot of knowledge which was used to refine further questioning and better extract relevant information which helped the iterative knowledge process which also shaped the scenarios (UNDP 2018).

FORESIGHT STRUCTURE OF THIS STUDY AND SCENARIO DEVELOPMENT

This foresight structure consists of the following components: a literature review-based horizon scanning, identification of weak signals, identification of wild cards, marking uncertainties, and constructing scenarios with the information gathered. Lastly, the scenarios will be presented to experts within the field to gather their opinions on how probable these scenarios are.

Horizon Scanning

The horizon scan is conducted as a continuation of the literature review conducted earlier. Scanning is not meant to predict what will occur, it acts as a tool to identify what might occur (UNDP 2018). To explore futures beyond the probable, this foresight study examines current trends and drivers, notably the growing integration of AI in both offensive and defensive cybersecurity and the increased use of the dark web for threat intelligence. In brief, the identified trend and drives from literature are:

1) *Offensive AI Applications*: AI systems capable of automating cyberattacks are reshaping the threat landscape, enabling malicious code generation, automated malware distribution, deepfakes, adversarial attacks, and botnet operations. 2) *Defensive AI Applications*: On the defensive side, organizations increasingly deploy AI-based systems to automate threat detection, vulnerability assessment, and incident response. Salem et al. (2024) highlight ongoing research into ML, DL, and metaheuristic approaches that help security teams analyze large datasets, detect attack patterns, and respond more rapidly to threats. 3)*Dark Web Intelligence*: A growing trend is the use of AI to monitor dark web activity for proactive threat intelligence. Kaur et al. (2023) highlight ongoing research using AI to analyze hacker forums and marketplaces through techniques such as sentiment analysis (IBM 2023) and topic modeling (Murel and Kavlakoglu 2024). Despite challenges from multilingual communication, approaches like bilingual lexical resources (BiSAL) improve analysis, enabling the identification of emerging threats, malicious actors, and Advance Persistent Threats (APTs) (Kaur et al. 2023; Cisco 2025).

Weak Signals

These elements act as indicators of change. Subtle shifts in trends or underlying patterns are often categorized as weak signals. While less visible or "noisy," they consist of raw information that can signal emerging change. Other indicators include drivers as the underlying forces that push change forward and provide critical context for understanding how and why future developments evolve in particular directions (UNDP 2018). The identified weak signals are:

1. *Regulatory Approaches to AI* Regulatory approaches to AI in cybersecurity are developing slowly and vary across countries. Mirsky et al. (2023) identify active efforts in the USA through the National Security Commission on Artificial Intelligence (NSCAI) (National Security Commission on Artificial Intelligence

(NSCAI) 2021) and the NIST AI Risk Management Framework (Tabassi 2023), the United Kingdom through the Principles for Security of Machine Learning (National Cyber Security Centre (NCSC, UK) 2022), and the European Union through the EU AI Act (European Parliament 2023) alongside the Cyber Resilience Act (European Commission, Digital Strategy 2024). Although these frameworks remain iterative and lack widespread adoption, they represent a strengthening weak signal toward future standardization.

2. *Integration of Threat Intelligence Sources*: Several documents highlight the need for effective threat intelligence platforms, yet implementation remains limited. Proper platforms would enable governments, security operators, and stakeholders to share timely and accurate threat information, strengthening collective defense (Kaur et al. 2023). This aligns with Malatji's call for international collaboration, as cybersecurity threats are global and require coordinated cross-border and cross-industry responses (Malatji 2023).

3. *Emergence of Highly Personalized Phishing attempts*: Offensive AI has enabled diverse attack vectors, with sophisticated phishing gaining particular traction. Attackers use AI-driven reconnaissance and synthetic media to craft highly personalized phishing campaigns (Mirsky et al. 2023). These attacks are especially effective because human error remains the weakest link, contributing to the rise in phishing incidents noted for example in Norway's NSM Risiko 2025 report (Nasjonal sikkerhetsmyndighet (NSM) 2025).

Wild Cards - Dangerous Signs

The events that have a low certainty of occurring, however if they would come to reality they might have catastrophic impact. Though wild cards may or may not be announced by weak signals, they often tend to mark the discovery of new opportunities or risks that should rise concerns. Wild card are categorized based on their visibility and impact: 1) *Black Swan*, refers to events of low probability, but high impact. 2) *Grey Swan*, refers to Predictable events to a certain extent. 3) *Red Swan*, refers to misleading signal that did not amount to anything. 4) *White Swan* refers to something people are aware of, but do not acknowledge it presence. The value of identifying wild cards in foresight lies in increasing preparedness by creating stressful scenarios to showcase vulnerabilities in proved "safe" environments. Wild cards were identified encompassing the low-probability, high-impact occurrences on cybersecurity landscape. They have been categorized into black swans and grey swans:

1. *AI-Quantum Computing Breakthrough (Black Swan)*: A major wild card is the emergence of practical quantum AI systems, which could undermine existing cryptography and destabilize current cybersecurity infrastructures (Brennan 2018). The devastating effects it would have on public key encryption would damage the integrity and confidentiality of digital communication and data storing platforms. NIST and other organizations have taken precautionary decisions and are working on "post-quantum" cryptography standards, however, a sudden breakthrough could create a massive security crisis event before any of these alternatives have been widely adopted.

2. *Autonomous AGI-Driven Cyber Attack (Black Swan)*: The emergence of fully autonomous Artificial General Intelligence (AGI) systems represents a major wild card. While AI already automates multiple stages of the cyber kill chain (Lockheed Martin 2023), AGI-enabled malware could autonomously adapt, learn, and conduct simultaneous multi-phase attacks across infrastructures (Guembe et al. 2022; Kazimierczak et al. 2024), posing an unprecedented cybersecurity threat.

3. *AI-Driven Cyber Warfare (Grey Swan)*: AI-driven warfare remains only partly predictable, as not all behaviors or threats can be fully modeled. AI is rapidly enhancing both offensive and defensive cyber capabilities, with cybercriminals and nation-states already weaponizing AI by exploiting weaknesses in ML models (Nobles 2024; Nobles 2023). As AI-driven attack tools become more accessible, even to less-skilled actors, offensive AI threats are increasingly plausible, particularly as many security teams remain largely human-driven and struggle to respond effectively (Guembe et al. 2022; Nobles 2023).

Uncertainties

Scenarios are developed around key uncertainties, as not all variables can be accounted for. These uncertainties are derived from horizon scanning and form the basis for scenario construction. Typically, two critical uncertainties are selected and mapped onto a 2×2 matrix, with each axis representing one uncertainty, providing a clear view of impacted areas and their implications (UNDP 2018).

While many uncertainties exist in AI-driven cybersecurity foresight, this study focuses on two: the technical sophistication of AI-based attack tools and the expertise of the attacker (See Figure 3). These uncertainties form the foundation of the scenario contexts.

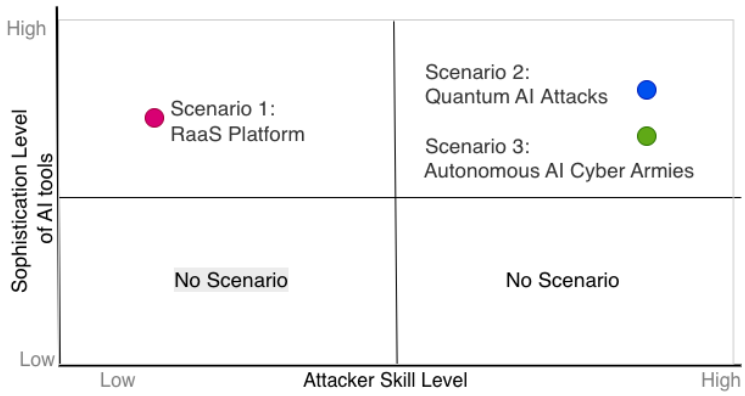

Figure 3. A 2x2 Matrix visualizing how the scenarios fit on the graph of uncertainties based on our assessment

1) *Level of Technical Sophistication*: This uncertainty reflects the rapid evolution of AI-based cyberattacks and attacker unpredictability. AI may range from a simple human-assisting tool to a fully autonomous cyber weapon, with its sophistication closely tied to attacker resources and operational scale.

2) *Level of Attacker Skill*: This uncertainty reflects how AI reshapes the threat landscape by lowering the skill barrier for cyberattacks. While even inexperienced attackers can now launch attacks with minimal effort, highly skilled hackers remain significantly more dangerous.

In Figure 3, the x-axis visualizes the skill of the attacker. The y-axis visualizes the technical sophistication of the AI tool they use. To strengthen the credibility of the scenario developments, key assumptions were defined, as presented in Table 1. These assumptions guided the identification of attackers, targets, scope, likelihood, and time-frame, addressing the question: *"Why is this scenario plausible, and under what conditions?"*.

Scenarios

The scenarios aim to be plausible and internally consistent, but do not assign probabilities or timelines due to inherent uncertainty. Each scenario includes a title, key uncertainties, contextual narrative, and implications, and is designed to provoke critical reflection and expert validation (UNDP 2018). Based on the literature review and horizon scan of AI-driven cyber threats, three scenarios were developed to represent emerging risks. These scenarios illustrate how AI enhances offensive capabilities and challenges security teams, using current knowledge to explore currently improbable futures and gather expert feedback on their plausibility.

Scenario 1: Ransomware-as-a-Service (RaaS). This scenario is classified as *High sophistication — Low attacker level*. The possible context is as follows: "A less skilled attacker purchases access to a RaaS platform to conduct cyberattacks using AI-powered tools: 1) *Conduct reconnaissance* using AI-driven OSINT (Open Source Intelligence) gathering to identify vulnerable targets. 2) *Launch spear-phishing campaigns* with AI-generated personalized emails to gain initial access. 3) *Lateral Movement:*Once inside a network, the AI ransomware adapts its behavior to avoid detection and finds optimal paths for lateral movement. 4) *Compromised:* The ransomware uses machine learning models to analyze the compromised data and set ransom demands on compromised assets. 4) *Adversarial AI:* If the victim tries to recover data, the ransomware uses adversarial AI techniques to avoid/defeat common recovery methods. **Scenario 1 — Background**: In today's society, services such as Cloud-as-a-Service (CaaS), Platform-as-a-Service (PaaS), Infrastructure-as-a-Service (IaaS), and Software-as-a-Service (SaaS) exist (Google Cloud n.d.); similarly, future ecosystems may give rise to Ransomware-as-a-Service (RaaS). This AI-based service would provide malicious actors with rentable tools for conducting AI-driven cyberattacks. Contrary to conventional assumptions, in this instance, this service does not require the attacker to possess excessive expertise to successfully utilize these tools. The malicious actor might possess limited knowledge on how the service works and how the cyberattack operation occurs necessarily. An attacker could use the service to perform OSINT-based reconnaissance, generate convincing spear-phishing campaigns using deepfakes or synthetic text, and deploy autonomous malware capable of lateral movement, data valuation, encryption using ransomware, and blocking user access. It also has the capabilities to block and repel basic self-defense mechanisms and delay the incident response process as it demands for a ransom to decrypt the targeted data.

Scenario 2: Quantum AI-Powered Cyberattacks. This scenario is classified as *High Technical Sophistication — High Attacker Skill*. The context can be that a government sponsored group of elite hackers achieves a

Table 1. Pre-defined Assumptions of Scenarios

Assumptions	Scenario 1	Scenario 2	Scenario 3
What AI capabilities exist in the scenario?	OSINT, Phishing, Ransomware	OSINT, Phishing, Sabotage, Malware, Quantum-computing	OSINT, Phishing, Sabotage, Malware, Autonomous AI, AGI
Are these tools general-purpose, custom-built, or available through underground services?	Available through dark-web for rental (underground services)	Custom-built, quantum-computer incorporating AI	Custom-built AGI
How autonomous or adaptive is the AI in carrying out attacks?	AI tools with human oversight.	AI tools with human oversight	Fully autonomous AI
What is the skill level and motivation of the attacker?	Can range from low to high, however, for the purpose of this scenario it is set to "Low".	High attacker skill, geopolitical motivations	High attacker skill, geopolitical and monetary motivations
Are they individuals, criminal groups, or state-sponsored actors?	Individuals.	State-sponsored group	Cybercriminal group
Do they rely on technical expertise or off-the-shelf AI tools?	Low-skill hackers relying on tools.	Highly capable hackers	Highly capable hackers
What systems or organizations are targeted?	Organization with weak cybersecurity fundamentals for easy monetary gain.	Enemy nation-states	Nation-states, wealthy organizations for monetary gain
Are defenses modern, legacy-based, or inconsistent?	Inconsistent.	Modern defenses	Varies from inconsistent to modern defenses
Does human error or exposure (e.g., public data) play a role?	Human error, main entry point is phishing.	Human error, main entry point is phishing.	Human error, main entry point is phishing.
What limits defenders' use of AI or other technologies?	Lack of resources, lack of staff	Defenders utilize state-of-the art AI	Varies from the defenders utilizing state-of-the-art AI to defenders lacking resources and staff to create a strong defense
Are policies or ethical constraints preventing an effective response?	x	Politics-infused	x
Is international coordination against such threats strong or weak?	Small scale hacking, no global cooperation mission is deployed.	Yes, this will create an AI arms race where states will invest	Yes, this will create an AI arms race where states will invest
When does the scenario take place (e.g., near future or far future)?	Near future.	Far future	Far future
What current trends support the plausibility of this development?	Low barrier of entry for AI tools, growing gap between offensive AI and defensive AI capabilities.	Developments of AGI, developments of quantum computers, developments of PQC (Post Quantum Cryptography).	Developments of AGI, Developments of autonomous AI

breakthrough in quantum computing, integrating it with specialized AI models to crack encryption and orchestrate devastating cyberattacks against global critical infrastructure: 1) *Break cryptography*: Attackers combine quantum

computers with advanced AI to develop tools that can instantly break modern cryptographic algorithms, decrypting communications and protected data. 2) *Targets*: The AI autonomously prioritizes targets (government, finance, energy) based on potential impact, using big data analysis and AI-driven OSINT. 3) *AI automation*: AI modules automate the deployment of malware and the exfiltration or manipulation of data, using quantum-powered attacks. 4) *Disruption*: The attackers simultaneously trigger disruptions; power grid failures, financial thefts and infrastructure sabotage now that most cryptography models are broken. **Scenario 2 — Background**: A central element of this scenario is the use of quantum computers to break today's encryption. AI-driven OSINT enables reconnaissance to identify high-value targets, after which quantum-powered AI systems can compromise encryption protocols previously considered secure, rendering current cryptography unreliable. Such capabilities are unlikely for individual hackers and more plausible for state-sponsored APTs with sufficient financial and computational resources. As a cyber warfare weapon, quantum-powered AI would fundamentally reshape the threat landscape, enabling cross-border operations for surveillance, disruption, or geopolitical advantage.

Scenario 3: Autonomous AI Cyber Armies. This scenario is classified as *High Technical Sophistication — High Attacker Skill*. Context: A highly skilled cybercriminal syndicate develops and unleashes a fully autonomous, self-directed AI system capable of orchestrating offensive cyber campaigns with minimal human oversight. This AI continually learns from its environment, adapts tactics in real time, and coordinates swarms of malware agents across the globe: 1) *Create AI Agent*: Attackers leverage advanced reinforcement learning and genetic algorithms to create an AI agent that can independently plan, launch, and refine cyberattacks. The AI scans the internet, autonomously identifying high-value targets and mapping vulnerabilities using AI-driven OSINT and anomaly detection. 2) *Scan Internet*: The AI scans the internet, autonomously identifying high-value targets and mapping vulnerabilities using AI-driven OSINT and anomaly detection. 3) *Attack Path*: The AI dynamically chooses and modifies attack paths, employing adversarial machine learning to evade detection, escalate privileges, and move laterally. 4) *Attack Path*: Upon gaining access, the AI independently assesses the most damaging or profitable action—data exfiltration, sabotage, ransomware, or disinformation—by analyzing system context and financial impact. **Scenario 3 — Background**: The emergence of fully autonomous AI-driven botnets and malware poses a major cybersecurity threat. In this scenario, an AGI-like system used by highly skilled actors operates without human oversight, learning continuously and coordinating attacks through swarm intelligence (Nobles 2024; Guembe et al. 2022). Capable of OSINT-based reconnaissance, real-time vulnerability exploitation, defense evasion, and data valuation, such AI-driven malware enables unprecedented speed, scale, and adaptability, marking a shift from human-led to AI-led cyber warfare.

Each scenario has also been illustrated in a cyberkill-chain mapping matrix (Lockheed Martin 2023) as a visualization tool. This is to help demonstrate how future threats may conduct an attack through every phase of the kill chain using AI technologies. The example of the scenario presentation using a cyberkill-chain approach can be seen in Figure 4.

FINDINGS

In this section, we will describe the empirical findings of this study. This section discusses the scenario interviews and the experts' perception of AI and the crafted scenarios.

Findings from Scenario Elicitation

Probability of Scenario 1: Ransomware-as-a-Service (RaaS)

After a general discussion and asking some specific questions about Future AI, we presented the first scenario to the interviewee and asked for their opinion on how likely they view the first scenario, could become a reality in the future based on what their knowledge about AI is, not only about cybersecurity specifically.

The expert feedback strongly suggests that Scenario 1 is not a distant or speculative future but rather an extension of attack patterns already observed today, amplified by AI. Several experts point out that many of the core elements of this scenario—long-term system infiltration, targeted social engineering, and financially motivated attacks—are already in use, albeit in more manual and fragmented forms. As AI-expert-1 explains, attackers today can remain inside organizational systems for extended periods and strike at precisely the right moment, such as when a company is about to transfer large sums of money. The expert describes familiar cases where attackers impersonate senior executives and issue convincing payment instructions, concluding that "this sounds likely" because the technique is already well established [AI-expert-1]. From this perspective, AI does not invent a new threat but refines timing, scale, and precision.

Other experts emphasize how rapidly AI development accelerates this evolution. AI-expert-2 notes that current progress in AI makes the scenario "highly possible," particularly the idea that ransomware could use machine

Threat Campaign	TC.01 - Small- to Mid-scale cyberattacks		
Threat Scenario	TS.01 - RaaS-platform used in cyber attacks		
Asset ID.	CA.SS.01	**Threat Actor ID.**	TA.E.01
Technical Sophistication	High	**Attacker Skill**	Low
Context			
A less skilled attacker purchases access to a RaaS platform to conduct cyber-attacks using AI-powered tools.			

Phase	Description
Reconnaissance	The AI tool provides means to conduct OSINT scans to find vulnerabilties in networks and users.
Weaponization	AI generates spear-phishing campaigns with text and deepfakes to mask payloads or malware.
Delivery	E-mail delivery of personalized phishing documents that contain malware.
Exploitation	The AI malware exploits the system vulnerabilities and weaknesses by utilizing evasion techniques upon gaining access.
Installation	The AI malware deploys backdoors and maneuvers around the system to find exploitable data.
Command & Control	The malware is able to connect to the cloud-based dashboard of the user to communicate via stealthly and encrypted channels.
Actions on Objectives	The AI compromises data and deploys ransomware for valuable data for financial gain, and disrupts the systems defense mechanisms with adversarial AI techniques.

Figure 4. Scenario Presentation Using Cyberkill Chain Approach

learning to analyze compromised data and automatically determine ransom demands. This capability resonates strongly with existing criminal incentives, as it optimizes financial gain while reducing human involvement [AI-expert-2]. The automation of judgment—deciding what data is valuable, how much a victim can pay, and when to apply pressure—marks a significant shift from human-led operations to AI-driven decision-making.

A recurring theme across the interviews is the erosion of skill barriers. AI-expert-3 highlights that AI introduces a form of reasoning into cyber operations, allowing even poorly constructed attacks to succeed. Unlike skilled attackers who understand consequences and limits, less experienced actors may act recklessly, yet AI compensates for their lack of strategy by providing adaptive and automated capabilities. As the expert notes, such attacks may be "reckless and poorly constructed," but they would "most probably" still work [AI-expert-3]. This raises concerns not only about sophistication but about volume—more attacks launched by more actors with fewer constraints.

AI-expert-4 extends this argument by focusing on accessibility and scale. The expert argues that AI-powered Ransomware-as-a-Service could turn "not that good hackers into better ones," and even enable people with no hacking knowledge to carry out direct attacks through services available on the dark web. The implication is a broad expansion of the threat actor landscape, where AI tools remove technical friction and enable misuse across criminal, terrorist, and even state-level contexts. As AI-expert-4 bluntly states, "I think it could very well happen," potentially even in warfare scenarios [AI-expert-4].

Taken together, the experts portray Scenario 1 as a highly plausible future shaped less by technological breakthroughs and more by the convergence of existing attack methods with AI-driven automation. The scenario reflects a shift in agency: from skilled individuals carefully crafting attacks to AI systems executing them at scale, with minimal oversight and increasing efficiency. This convergence explains why all experts ultimately assess the probability of this scenario as high—not because it represents a radical leap, but because it builds naturally on trends already unfolding today.

Scenario 2: Quantum AI-Powered Cyberattacks

Following the first scenario, we introduce the second scenario to the experts, which also has been described in Scenario section earlier.

Scenario 2 evokes a more unsettling and less immediately visible future, one shaped by breakthroughs that may already be unfolding beyond public view. While the experts agree that quantum AI–powered cyberattacks are technically plausible, their assessments reflect caution, uncertainty, and a shared sense that this scenario belongs to a different category of threat—one defined by scale, power, and geopolitics rather than everyday cybercrime.

AI-expert-1 views this scenario as likely, largely because of the pace of advances in both quantum computing and cryptography research. The expert points out that while post-quantum encryption is actively being developed, the threat does not only concern future data. Instead, there is a strong belief that powerful states are already "storing very large amounts of data" with the expectation that it can be decrypted later, once quantum capabilities mature [AI-expert-1]. This "decrypt later" possibility gives the scenario a chilling dimension: even historical communications and sensitive records may become vulnerable retroactively. The expert further highlights the risk of long-term, undetected access to critical systems, where attackers could remain dormant until a strategic moment arises—describing this prospect as "very scary" and extremely difficult to detect [AI-expert-1]. Thus, in relations long-term data governance strategies prioritizing post-quantum migration for high retention records and segmenting stored data to minimize risks.

AI-expert-2 also acknowledges the feasibility of such attacks, particularly against critical infrastructure, but frames the scenario as highly dependent on the speed at which quantum computing and AI converge. If this convergence accelerates, the expert believes such attacks could emerge within a relatively short timeframe, potentially within five years [AI-expert-2]. This assessment reflects a conditional outlook: the threat is real, but its timing hinges on technological breakthroughs that remain uncertain.

In contrast, AI-expert-3 introduces a more restrained perspective, arguing that while quantum AI may exist, it is unlikely to be used casually. The expert characterizes it as a strategic weapon rather than a tool for routine cyber operations, suggesting that its deployment would likely be reserved for wartime or extreme geopolitical conflict. As the expert explains, many countries may possess such capabilities, but using them—especially against power grids or critical infrastructure—would constitute an act of war, making restraint more likely than frequent use [AI-expert-3].

AI-expert-4 reinforces this distinction by emphasizing access and scale. While acknowledging the possibility of quantum AI being used in cyber warfare, the expert argues that it is "less likely" than Scenario 1, primarily because such technology would be limited to governments or large, well-funded organizations. Unlike AI-driven ransomware services, quantum AI would not be widely accessible, nor deployed at scale by ordinary threat actors [AI-expert-4].

Taken together, the expert insights frame Scenario 2 as a high-impact but lower-probability future—one that carries profound consequences if realized but is constrained by cost, complexity, and geopolitical risk. Unlike Scenario 1, which evolves organically from today's threat landscape, Scenario 2 represents a strategic inflection point, where cyber operations begin to resemble weapons of deterrence rather than tools of crime. The uncertainty does not lie in whether quantum AI could change cybersecurity, but in whether—and under what conditions—it would be unleashed.

Scenario 3: Autonomous AI Cyber Armies

The third and final scenario, provoked the following opinions from the interviewees:

Scenario 3 elicited a mix of concern, caution, and conditional acceptance from the experts, reflecting both the disruptive potential and the uncertainty surrounding fully autonomous AI-driven cyber operations. AI-expert-1 reacts viscerally, describing the scenario as "very scary," yet expresses hope that such capabilities remain far in the future, emphasizing that current technology is still "too advanced" for this level of autonomy to be fully realized [AI-expert-1]. This sense of distance, however, is not shared uniformly. AI-expert-2 argues that many of the building blocks for this scenario already exist, particularly in areas such as reinforcement learning, multi-agent systems, and autonomous decision-making. They stress that AI independently assessing "the most damaging or profitable actions" is not speculative but actively researched, calling the scenario "very dangerous and worrisome" and suggesting that breakthroughs could arrive in the near future [AI-expert-2]. The expert further warns that attackers often outpace defenders, reinforcing the need for governments and large institutions—such as the European Commission—to begin planning now rather than reacting later.

AI-expert-3 places Scenario 3 between Scenarios 1 and 2 in terms of likelihood, arguing that its realization depends largely on the degree of autonomy achieved. From this perspective, attackers are drawn to tools that maximize

automation and convenience, and AI fits naturally into this logic. The expert likens using AI-driven attack tools to "driving a car," noting that one does not need deep technical understanding to operate them effectively, making such systems "low-hanging fruit" for malicious actors [AI-expert-3]. This accessibility, combined with efficiency gains, makes widespread adoption plausible once the tools mature.

AI-expert-4 echoes this ambivalence, invoking cultural references such as "Terminator and Skynet" to illustrate the negative connotations surrounding autonomous AI, particularly concerns about accountability and opaque decision-making in black-box systems [AI-expert-4]. While acknowledging the potential dangers of losing human oversight in cyberattacks, the expert also notes that similar technologies could be used defensively, for detection or mitigation. Ultimately, they assess the scenario as less likely than Scenario 1 but more plausible than Scenario 2, reflecting a middle-ground position shared by several experts.

Taken together, Scenario 3 is framed as a transitional future: not yet imminent, but increasingly plausible as autonomy, learning, and coordination capabilities advance. The experts converge on the idea that while fully autonomous AI cyber armies may not dominate today's threat landscape, the trajectory of AI research—combined with attackers' preference for automation—makes this scenario one that defenders cannot afford to ignore.

Scenario assessment table

Based on the feedback from each individual interviewee, a table to display their trust that a scenario will sometime in the future become a reality was created to visualize the responses. Note that the topics of the interviews were heavily skewed based on the area of expertise of each interviewee. Thus, some experts went deeper into certain topics than others, and their opinion on the plausibility of the scenarios also varied. Their responses have been standardized to display coherence and comparability. The table shows three possibilities, **High, Medium, Low**, where *High* means that the scenario has a high probability of occurrence, *Medium* means that the interviewee had trust in the event happening sometime, and *Low* infers that the interviewee was skeptical in the scenario occurring and is a somewhat futuristic or implausible event.

Table 2. Probability of scenario assessment table based on interview feedback

	Expert 1	**Expert 2**	**Expert 3**	**Expert 4**
Scenario 1	High	High	High	High
Scenario 2	High	Medium	Low	Low
Scenario 3	Low	Medium	Medium	Medium

Findings from Interviews on AI-driven Cybersecurity

The variations of expert judgments can also be affected by their individual background. For instance, judgment from the expert with AI-enabled crisis detection and misinformation analysis background may use social media and open-source data. This background can influence scenario appraisal by prioritizing publicly observable signals and socio-technical threat mechanisms, while potentially assigning lower likelihood to stealthy or highly technical cyber pathways that are less visible in their typical datasets. These differences can be interpreted as a lens effect rather than an error, and they add diversity to the assessment.

Beyond the scenario-specific discussions, participants also shared broad reflections on artificial intelligence, particularly regarding the future capabilities, risks, and societal impacts. This section summarizes key themes from this part of the interviews, highlighting attitudes that inform the broader context of AI-driven cybersecurity.

Perceived Capabilities and Limits of AI

Across the interviews, AI is consistently described not as a traditional, rule-based tool but as a system defined by learning, adaptation, and uncertainty. AI-expert-1 highlights this distinction clearly, explaining that modern AI is "not pre-programmed" but instead trained on data to make predictions in complex, unfamiliar situations [AI-expert-1]. This ability to generalize from past data is seen as AI's core strength, particularly in environments like cybersecurity where conditions change rapidly. However, this same dependence on data also introduces fragility. Several experts express concern that AI systems may increasingly rely on reused or outdated data, which could lead to hallucinations or incorrect decisions—an especially serious issue in high-stakes domains such as threat detection and authentication [AI-expert-1]. As large models exhaust publicly available text, future training may shift toward audio and video data, raising new privacy and surveillance concerns alongside technical challenges.

Quantum Computing, Power, and Uncertainty

The convergence of AI and quantum computing is widely viewed as both transformative and unsettling. AI-expert-1 bluntly associates quantum computing with its potential to "break encryption," reflecting a common fear within cybersecurity communities that current cryptographic systems may become obsolete [AI-expert-1]. AI-expert-2 expands on this concern, describing the field as highly uncertain but potentially "very problematic for humanity" if quantum-AI capabilities mature under the control of only a few powerful actors [AI-expert-2]. While progress remains largely theoretical and uneven across countries, experts agree that the stakes are high. AI-expert-3 takes a more cautious stance, noting that quantum computing may excel only in specific niche problems rather than general-purpose applications [AI-expert-3]. AI-expert-4 frames the issue as an arms race, emphasizing that the impact of quantum AI depends entirely on who controls it—whether it is used to protect society or to dominate it [AI-expert-4]. Across perspectives, access, regulation, and preparedness emerge as more critical than raw technical capability.

Vulnerabilities, Human Weakness, and the Role of AI in Defense

Despite AI's strengths, experts repeatedly stress that AI systems do not truly understand their actions and remain vulnerable to manipulation. AI-expert-1 explains that AI performs well within its training boundaries but can be "tricked" by small, often invisible perturbations, particularly in areas like image recognition and facial authentication [AI-expert-1]. These weaknesses mirror human vulnerabilities rather than replace them. Humans, while prone to phishing and deception, often retain contextual awareness and intuition—questioning whether a message "makes sense" in a given moment [AI-expert-1]. This reinforces a recurring theme: AI and humans should complement rather than replace one another.

On the defensive side, AI is widely seen as valuable for detecting anomalies, scanning large datasets, and identifying unusual patterns that humans might overlook. AI-expert-1 describes how AI can flag suspicious communications or unfamiliar activity across systems [AI-expert-1], while AI-expert-3 highlights its usefulness in proactively identifying weaknesses in IoT devices, such as poor passwords or exposed ports [AI-expert-3]. However, this defensive advantage is constantly challenged. As AI-expert-3 and AI-expert-4 both note, advances in deepfakes and synthetic media fuel an ongoing cycle in which attackers improve deception while defenders improve detection—an arms race with no final solution [AI-expert-3][AI-expert-4].

Autonomy, Explainability, and Trust

The idea of fully autonomous AI agents provokes both optimism and unease. AI-expert-1 cautiously supports the use of AI agents within tightly defined limits, warning that autonomy without constraints requires "utmost carefulness" [AI-expert-1]. In contrast, AI-expert-2 emphasizes how autonomy enables increasingly dangerous applications, from adaptive malware to large-scale disinformation campaigns that evolve in real time [AI-expert-2]. AI-expert-4 brings the discussion back to trust, questioning whether organizations and societies will accept systems whose decisions cannot be fully explained, especially in critical domains [AI-expert-4].

Explainability emerges as a central unresolved challenge. Experts agree that the most powerful AI models—deep neural networks and transformers—are also the least transparent. As AI-expert-1 and AI-expert-2 note, this opacity makes it difficult to verify decisions, identify failures, or assign accountability, which limits how far automation can realistically go in cybersecurity [AI-expert-1] and [AI-expert-2]. As a result, most experts converge on the same conclusion: fully autonomous AI is neither realistic nor desirable in the near term. Instead, a human–AI symbiosis, where AI augments human judgment rather than replaces it, is seen as the most viable path forward.

Acceleration and the Emerging AI Race

Looking ahead, AI-expert-2 summarizes the broader trajectory, describing the coming decade as an "AI race" in cybersecurity—one driven equally by attackers and defenders [AI-expert-2]. This framing captures the underlying tension present throughout the interviews: AI accelerates both protection and harm, innovation and exploitation. The challenge is no longer whether AI will shape cybersecurity, but who will shape AI, under what rules, and to whose advantage.

DISCUSSION AND IMPLICATIONS

Discussion

This study explores how AI is currently used in cybersecurity, both to enable cyberattacks and to strengthen cyber defense, with the aim of helping present and future professionals navigate an increasingly complex threat

landscape. By combining primary interview data with insights from the literature, the analysis highlights not only the opportunities AI brings to cybersecurity but also the new vulnerabilities it introduces. The discussion reflects on the evolving role of human actors, the technical and ethical challenges of AI development, and how the foresight methodology contributed to a deeper understanding of future AI-related risks. Expert validation plays a central role in assessing the plausibility of the scenarios developed in this foresight study. Although the empirical data is based on four interviews, the processes were iterative and the scenarios were firmly grounded in the literature, which strengthens their credibility even as their timelines remain uncertain. The expert assessments and contextualized interview results provide insight into how different futures may unfold and why certain developments appear more likely than others.

There was strong consensus among the experts that the first scenario—an AI-enabled Ransomware-as-a-Service platform—is highly probable. This assessment closely aligns with existing research showing that AI lowers the barrier to entry for cybercrime by making powerful tools accessible to less skilled actors. In this scenario, cyberattacks become more industrialized, scalable, and efficient, enabling high-impact operations with minimal technical expertise. Several experts also noted that such platforms could extend beyond financially motivated crime and be leveraged in warfare contexts by states or terrorist organizations. From the authors' perspective, the widespread adoption of cloud-based service models makes it realistic that offensive AI capabilities could be offered as rentable services, reinforcing the high probability assigned to this scenario.

The second scenario—quantum AI-powered cyberattacks—elicited more uncertainty. Expert assessments ranged from high to low probability, largely reflecting limited familiarity with the convergence of quantum computing and AI and the early stage of both academic and applied research in this area. This uncertainty is also evident in the literature, where quantum computing plays a relatively minor role compared to more immediate AI-driven threats. Still, a shared concern emerged around cryptography, with one expert emphasizing that "the biggest driving force behind quantum computing development is for breaking encryption." This aligns directly with the scenario's focus on exploiting cryptographic vulnerabilities. While ongoing research into post-quantum cryptography suggests growing awareness of this risk, the authors conclude that the timing of such attacks remains highly uncertain and therefore assess this scenario as having a medium probability.

The third scenario, involving fully autonomous AI cyber armies, generated cautious and divided responses. Experts generally viewed this future as less likely than the first scenario but potentially more plausible than the quantum AI scenario. Some pointed to rapid progress in reinforcement learning, multi-agent systems, and automation, suggesting that increasingly autonomous cyber tools are technically feasible. At the same time, both expert opinion and the literature emphasize that AI systems are unlikely to operate fully autonomously without human oversight in the foreseeable future. Rather, AI is expected to function as part of a human–AI collaboration model, enabling augmented intelligence rather than independent agency (Nobles 2023). While the authors acknowledge that AI may achieve autonomy in specific, bounded tasks, the vision of self-directed systems that learn, adapt, and conduct complex cyber operations end-to-end is considered improbable in the near term, leading to a low probability assessment.

Overall, the scenario analysis illustrates a future shaped less by a single disruptive breakthrough and more by uneven progress across multiple dimensions of AI development. Some threats, such as AI-enabled cybercrime services, appear imminent and scalable, while others remain constrained by technical, ethical, and organizational barriers. The foresight approach proves valuable in capturing this uncertainty, allowing cybersecurity professionals to prepare not for one predicted future, but for a range of plausible trajectories driven by AI's continued integration into cyber operations.

Implications

The findings of this foresight study can be understood as comprising two broad temporal horizons: *developments that are already unfolding or likely to materialize in the near future*, and *more uncertain futures that emerge from weak signals and wild cards* derived from the literature and empirical data. Indeed, the precise timing and likelihood of these developments remain unclear, they provide a structured view of how AI is transforming the cybersecurity landscape.

In the near term, the study shows that AI-driven threats are not speculative but already active and evolving. The systematic literature review conducted in this study reveals already a wide range of offensive AI applications, captured by the diversity of themes in Figure 5.

Despite this variety, expert interviews consistently point to social engineering as the most dominant and effective attack vector. AI has significantly amplified this threat by enabling attackers to automate reconnaissance through OSINT and to generate highly convincing spear-phishing campaigns using natural language processing and

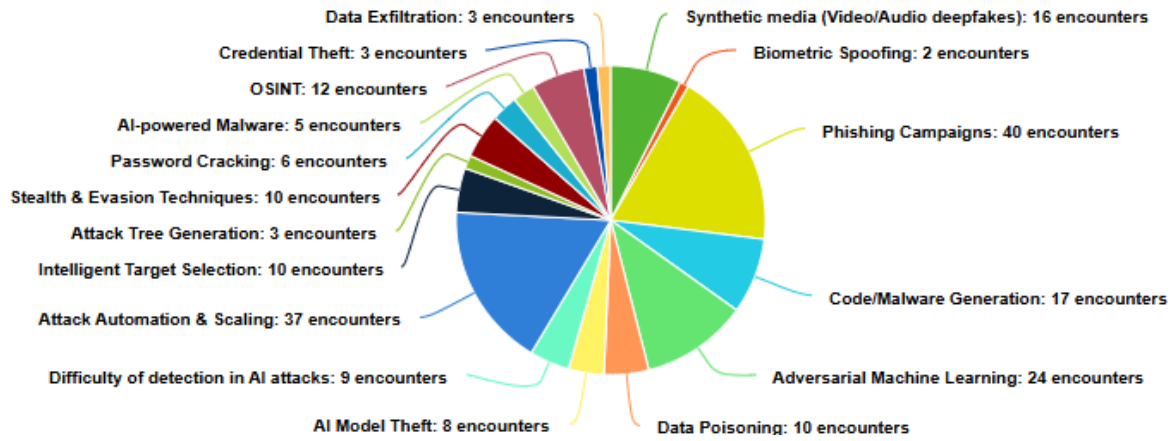

Figure 5. The themes identified for Offensive AI use through the literature review.

deepfakes. This finding aligns with broader institutional report such as ENISA's *Threat Landscape 2024*, which diagnoses phishing as the prominent cause of cybersecurity incidents, accounting for 73 percent of all social engineering attacks (ENISA 2024b, p. 63). As one expert noted, *"Humans have a substantial likelihood of errors and can easily be fooled by phishing techniques and alike,"* underscoring why people remain the primary targets. With AI becoming cheaper and more accessible, attackers can now extent these operations dramatically, automating data distillation, personalization, and mass deployment of attacks (Mirsky et al. 2023). This has unfolded multiple phishing avenues, ranging from *clone phishing* and *whaling* to *smishing, vishing,* and social-media-based attacks, confirmed by expert observations on how online platforms are used to spread disinformation and manipulation.

On the other hand, defensive efforts are racing to keep pace. Researchers and industry actors are developing AI-based detection tools capable of identifying synthetic text, images, video, and audio, and watermarking AI-generated content to limit misuse, such as Google's SynthID (Google 2023). As one expert pointed out, *"We do have such tools to detect plagiarism and also to detect AI,"* underlining that defensive innovation is active but reactive. This dynamic feeds into a broader asymmetry identified throughout the literature and validated by the experts: offensive AI capabilities are advancing faster and benefiting from greater incentives than defensive ones. While defensive AI excels at processing large volumes of data for threat detection, intrusion monitoring, and automated response (Nobles 2023; Truong et al. 2020), it is highly dependent on high-quality, up-to-date training data. Defensive models are vulnerable to data poisoning and struggle with unknown attack patterns (Malatji 2023; Kazimierczak et al. 2024). Meanwhile, attackers enjoy the advantage of initiative, innovation, and resource flexibility. This imbalance is further exacerbated by a shortage of cybersecurity professionals with deep AI expertise and the lack of shared platforms for coordinated threat intelligence, widening the gap between attackers and defenders. As one expert noted in relation to Ransomware-as-a-Service, AI can *"[...] make not that good hackers into better ones by supporting them."*

A recurring near-term challenge is trust. Both literature and interviews emphasize that the black-box nature of modern AI systems remains a critical weakness in cybersecurity contexts (Kumar et al. 2024). Deep neural networks and transformer-based models offer strong performance but little transparency, making it difficult for security teams to understand, audit, or justify AI-driven decisions. As one expert explained, *"AI models are generally not transparent and the humans have little control over it,"*, while also noting that more explainable models often lack the sophistication needed for complex tasks. This lack of explainability undermines trust, complicates incident response, and increases susceptibility to adversarial manipulation. As a result, most experts and sources converge on the need for human oversight and hybrid approaches, where AI augments rather than replaces human judgment (Nobles 2023). Research into explainable AI (xAI) is frequently proposed as a partial solution,to make AI decisions more transparent and verifiable (Kaur et al. 2023).

Beyond these near-term dynamics, the foresight analysis highlights more uncertain but potentially disruptive futures. One such development is the emergence of AI-quantum computing, framed as a black swan scenario due to its potentially catastrophic impact on modern encryption standards (Brennan 2018). The chance of states invest massively in quantum-powered AI and lead to quantum arms race for strategic advantage, has raises global cybersecurity concerns. Although precautionary measures are emerging, such as post-quantum cryptographic standards from NIST (2024), expert feedback suggests that quantum-powered AI remains a plausible future risk, even if its timeline is uncertain.

Another uncertain trajectory concerns AI autonomy. While current AI systems remain largely task-specific and dependent on human input (Mirsky et al. 2023), experts observe a clear trend toward autonomous agents—systems designed to act independently within defined boundaries. As one expert remarked, "many things within AI now is going in the direction of what is called agents." If combined with advances toward Artificial General Intelligence, such systems could operate across multiple stages of the cyber kill chain in parallel (Lockheed Martin 2023), dramatically increasing attack speed and scale. However, at present, fully autonomous AI capable of complex, end-to-end cyber operations remains unlikely, as most AI remains narrow and context-bound (Nobles 2023).

These developments culminate in the broader concern of AI-driven cyber warfare. Although some aspects of this future can be anticipated through current trends, the exact trajectory remains uncertain. As one expert succinctly stated, *"I think the next decade will be the AI race."* AI is increasingly positioned as a strategic asset in both offense and defense, reshaping geopolitical dynamics by enabling conflict below the threshold of conventional warfare. Regulatory responses are emerging slowly, creating gaps that malicious actors or state-sponsored programs may exploit. AI itself may assist in generating governance and compliance frameworks more efficiently (McIntosh et al. 2023), these tools remain vulnerable to manipulation and lack transparency. Ongoing efforts by organizations such as ENISA and NSCAI to regulate and secure AI systems signal progress, but slow adoption and standardization risk creating a regulatory vacuum. As one expert cautioned, *"It can be smart to assume that quantum and AI development have come further than what we can see publicly."*

To sum up, the findings highlight a future marked by accelerating asymmetry, increasing automation, and growing uncertainty. Offensive incentives currently outpace defensive ones, and without coordinated international collaboration, information sharing, and investment in explainable and resilient AI systems, this imbalance is likely to deepen. Our foresight study shows that encountering AI-driven cyber threats requires technical solutions, and strategic coordination across borders, industries, and disciplines to treat cybersecurity as a crucial global concern.

However, while foresight studies help exploring possible futures, but they have limits: they do not predict what will happen, and the further into the future they look, the more uncertainty there is. Results can be shaped by the people involved and their assumptions, which can introduce bias. Evidence is often limited for emerging topics, so the analysis may rely on expert judgment rather than strong data. Different methods can also produce different outcomes, and the findings can be hard to turn into clear, practical decisions. Thus, conducting a systematic literature review strengthens the empirical basis of the foresight study by making assumptions explicit, limiting reliance on subjective judgments, and improving transparency; however, it cannot remove fundamental uncertainty.

CONCLUSIONS

In an era where AI is deeply embedded in digital infrastructures, cybersecurity has become both a technical and societal challenge. This article set out to examine how AI reshapes cyber threats and defenses, and how foresight can be used to better prepare for what lies ahead. Instead of treating the future as a fixed outcome, this study emphasized the importance of proactive thinking—asking not only what threats may emerge, but how cybersecurity professionals can actively shape more resilient defenses in response to AI-driven risks.

To address the research question: *how foresight analysis of future AI capabilities can inform proactive cybersecurity strategies*, this article deemed the usefulness of the strategic foresight methodology grounded in literature and expert insight. It enables this study uncovering current AI-enabled threats, such as social engineering and adversarial attacks, and uncertain elements but potentially disruptive developments, including quantum-powered AI, autonomous systems, and AI-driven warfare. The scenario validation process showed variations among experts when judged the likelihood and timing. However, none of the scenarios were deemed as implausible, strengthening the value of foresight in navigating uncertainty rather than predicting exact outcomes. Moreover, adding more type of experts such as malware and digital forensic specialists in the scenario validation can strengthen the credibility of scenarios.

Overall, the path to our findings illustrates that foresight analysis offers a practical framework for enhancing cybersecurity preparedness by revealing asymmetries between offensive and defensive AI, highlighting the continued importance of human oversight, and encouraging early precautionary action. As AI capabilities continue to evolve, the key contribution of this work lies in showing that cybersecurity resilience depends not only on technical solutions, but also on anticipatory thinking, cross-sector collaboration, and the integration of human judgment with AI-driven tools.

REFERENCES

AIS (n.d.). *Research - Senior Scholars' List of Premier Journals*.

Arp, D., Quiring, E., Pendlebury, F., Warnecke, A., Pierazzi, F., Wressnegger, C., Cavallaro, L., and Rieck, K. (2021). *Dos and Don'ts of Machine Learning in Computer Security*. arXiv: 2010.09470 [cs.CR].

Brennan, D. (2018). "Quantum Computational Supremacy: Security and Vulnerability in a New Paradigm". In: *Irish Communication Review* 16, p. 10.

Castagnaro, A., Conti, M., and Pajola, L. (2024). *Offensive AI: Enhancing Directory Brute-forcing Attack with the Use of Language Models*. arXiv: 2404.14138 [cs.CR].

Cisco (Apr. 14, 2025). *What Is an Advanced Persistent Threat (APT)?* Cisco. URL: https://www.cisco.com/site/us/en/learn/topics/security/what-is-an-advanced-persistent-threat-apt.html.

ENISA (Mar. 2024a). *ENISA Cybersecurity Threats Foresight 2030: Update*. Technical Report. Belgium: Publications Office of the European Union.

ENISA (Sept. 19, 2024b). *ENISA Threat Landscape 2024*. Report. European Union Agency for Cybersecurity.

European Commission, Digital Strategy (2024). *Cyber Resilience Act*. URL: https://digital-strategy.ec.europa.eu/en/policies/cyber-resilience-act.

European Parliament (June 2023). *EU AI Act: First Regulation on Artificial Intelligence*. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

Giaoutzi, M. and Sapio, B., eds. (2013). *Recent Developments in Foresight Methodologies*. 1st ed. Vol. 1. Complex Networks and Dynamic Systems. New York, NY: Springer New York, pp. XVIII, 310.

Google (2023). *Identifying AI-generated images with SynthID*. Google DeepMind. URL: https://deepmind.google/science/synthid/.

Google Cloud (n.d.). *PaaS vs IaaS vs SaaS*. URL: https://cloud.google.com/learn/paas-vs-iaas-vs-saas.

Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., and and, V. P. (2022). "The Emerging Threat of Ai-driven Cyber Attacks: A Review". In: *Applied Artificial Intelligence* 36.1, p. 2037254. eprint: https://doi.org/10.1080/08839514.2022.2037254.

IBM (Aug. 24, 2023). *What Is Sentiment Analysis?* IBM. URL: https://www.ibm.com/think/topics/sentiment-analysis.

Information security, cybersecurity and privacy protection — Guidance on managing information security risks (Oct. 2022). International Standard. Geneva, Switzerland: International Organization for Standardization; International Electrotechnical Commission.

Iturbe, E., Llorente-Vazquez, O., Rego, A., Rios, E., and Toledo, N. (2024). "Unleashing offensive artificial intelligence: Automated attack technique code generation". In: *Computers & Security* 147, p. 104077.

Jacobsen, B. and Hirvensalo, I. (Nov. 2019). *9 Foresight Methodologies Successful Companies Use to Stay Ahead*. Futures Platform. URL: https://www.futuresplatform.com/blog/9-foresight-methodologies-successful-companies-use-stay-ahead.

Kamoun, F., Iqbal, F., Esseghir, M. A., and Baker, T. (2020). "AI and machine learning: A mixed blessing for cybersecurity". In: *2020 International Symposium on Networks, Computers and Communications (ISNCC)*, pp. 1–7.

Kaur, R., Gabrijelčič, D., and Klobučar, T. (2023). "Artificial intelligence for cybersecurity: Literature review and future research directions". In: *Information Fusion* 97, p. 101804.

Kazimierczak, M., Habib, N., Chan, J. H., and Thanapattheerakul, T. (2024). "Impact of AI on the Cyber Kill Chain: A Systematic Review". In: *Heliyon* 10.24, e40699.

Kumar, S., Datta, S., Singh, V., Singh, S. K., and Sharma, R. (2024). "Opportunities and Challenges in Data-Centric AI". In: *IEEE Access* 12, pp. 33173–33189.

Lockheed Martin (2023). *Cyber Kill Chain®*. URL: https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html.

Machado, C. and Medeiros Fröhlich, A. A. (2018). "IoT Data Integrity Verification for Cyber-Physical Systems Using Blockchain". In: *2018 IEEE 21st International Symposium on Real-Time Distributed Computing (ISORC)*, pp. 83–90.

Malatji, M. (2023). "Offensive Artificial Intelligence: Current State of the Art and Future Directions". In: *2023 - International Conference on Digital Applications, Transformation & Economy (ICDATE)*, pp. 1–6.

McIntosh, T., Liu, T., Susnjak, T., Alavizadeh, H., Ng, A., Nowrozy, R., and Watters, P. (2023). "Harnessing GPT-4 for generation of cybersecurity GRC policies: A focus on ransomware attack mitigation". In: *Computers & Security* 134, p. 103424.

Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, W., Elovici, Y., et al. (2023). "The Threat of Offensive AI to Organizations". In: *Computers & Security* 124, p. 103006.

Murel, J. and Kavlakoglu, E. (Mar. 30, 2024). *What Is Topic Modeling?* IBM. URL: https://www.ibm.com/think/topics/topic-modeling.

Nasjonal sikkerhetsmyndighet (NSM) (Feb. 2025). *Risiko 2025*. Annual Risk Assessment Report. Sandvika, Norway: Nasjonal sikkerhetsmyndighet.

National Cyber Security Centre (NCSC, UK) (2022). *Principles for the Security of Machine Learning*. URL: https://www.ncsc.gov.uk/collection/machine-learning-principles.

National Institute of Standards and Technology (Aug. 2024). *NIST Releases First 3 Finalized Post-Quantum Encryption Standards*. NIST. URL: https://www.nist.gov/news-events/news/2024/08/nist-releases-first-3-finalized-post-quantum-encryption-standards.

National Security Commission on Artificial Intelligence (NSCAI) (2021). *Final Report*. online.

Nobles, C. (June 2023). "Offensive Artificial Intelligence in Cybersecurity: Techniques, Challenges, and Ethical Considerations". In: pp. 348–363.

Nobles, C. (2024). "The Weaponization of Artificial Intelligence in Cybersecurity: A Systematic Review". In: *Procedia Computer Science* 239, pp. 547–555.

Norwegian Directorate for Higher Education and Skills (Feb. 2024). *The Norwegian Register for Scientific Journals, Series and Publishers (Kanalregisteret)*. Direktoratet for høyre utdanning og kompetanse. URL: https://kanalregister.hkdir.no/.

Page, M., Mckenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., Shamseer, L., Tetzlaff, J., Akl, E., Brennan, S., et al. (Mar. 2021). "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews". In: *BMJ* 372, n71.

Popper, R. (Oct. 2008). "How are foresight methods selected?" In: *Foresight* 10.6, pp. 62–89.

Poremba, S. (Jan. 15, 2025). *ISC2 Cybersecurity Workforce Study: Shortage of AI skilled workers*. IBM Think. URL: https://www.ibm.com/think/insights/isc2-cybersecurity-workforce-study-shortage-ai-skilled-workers.

Radianti, J. (2024). "Vision for Emergency Management: Conceptualizing Mission-Critical Ecosystems". In: *Proceedings of the International ISCRAM Conference*.

Radianti, J. (2025). "Navigating Digital Resilience in Complex Emergency Management Environments". In: *Proceedings of the International ISCRAM Conference*.

Salem, A. H., Azzam, S. M., Emam, O. E., and Abohany, A. A. (2024). "Advancing cybersecurity: a comprehensive review of AI-driven detection techniques". In: *Journal of Big Data* 11.1, p. 105.

Schlagwein, D., Currie, W., Leimeister, J. M., and Willcocks, L. (Jan. 2025). "Digital futures: Definition (what), importance (why) and methods (how)". In: *Journal of Information Technology* 40.1, pp. 2–8.

Schmitt, M. and Flechais, I. (2023). *Digital Deception: Generative Artificial Intelligence in Social Engineering and Phishing*. arXiv: 2310.13715 [cs.CR].

Sen, R., Heim, G., and Zhu, Q. (Jan. 2022). "Artificial Intelligence and Machine Learning in Cybersecurity: Applications, Challenges, and Opportunities for MIS Academics". In: *Communications of the Association for Information Systems* 51, pp. 179–209.

Tabassi, E. (Jan. 2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology.

Truong, T. C., Zelinka, I., Plucar, J., Čandík, M., and Šulc, V. (2020). "Artificial Intelligence and Cybersecurity: Past, Presence, and Future". In: *Artificial Intelligence and Evolutionary Computations in Engineering Systems*. Ed. by S. S. Dash, C. Lakshmi, S. Das, and B. K. Panigrahi. Singapore: Springer Singapore, pp. 351–363.

UK National Cyber Security Centre (2023). *Threat Modelling*. URL: https://www.ncsc.gov.uk/collection/risk-management/threat-modelling.

UNDP (2018). *Foresight Manual: Empowered Futures for the 2030 Agenda*. Singapore: United Nations Development Programme Global Centre for Public Service Excellence.

Zakaria, J. and Whitfield, B. (Jan. 2025). *Data Standardization: How to Do It and Why It Matters*. URL: https://builtin.com/data-science/when-and-why-standardize-your-data.