DS Journal of Artificial Intelligence and Robotics
Volume 3 Issue 4, 1-43, Oct - Dec 2025
ISSN: 2583-9926 / https://doi.org/10.59232/AIR-V3I4P101

Review Article

A Systematic Review of Artificial Intelligence (AI)-Enabled Cybercrime and Self-Directed Attack Systems

Taban Habibu[1,2*], Tonny Odoch[1*], Ceaser Obudra[1], Deograticus Afimani[1], Boniface Kadabara[1], Francis Xavier Ovoni[1], Benard Kasule[1]

[1]*Department of Computer and Information Science, Faculty of Technoscience, Muni University, P.O. Box 725, Arua, Uganda.*
[2]*Department of Computer Science, Islamic University in Uganda, Faculty of Science, P.O. Box 7244, Kampala, Uganda.*

[2]sultannubi@gmail.com

Received: 05 October 2025; Revised: 16 November 2025; Accepted: 08 December 2025; Published: 25 December 2025

Abstract - AI has been rapidly transforming how cybercriminals operate and how they attack. The use of AI will greatly affect the scope and level of sophistication that cybercriminals can achieve while executing attacks against organizations. The present work focuses on the development and growth of autonomous systems of attack that are developed and built around AI technology, the ability of AI to learn, reason, and adapt independently, and how the use of AI by cybercriminals can enable them to utilize AI at every stage of the cyber kill chain – Reconnaissance to Command and Control. Through an examination of 46 published studies from 2020 to 2025, the current analysis illustrates the speed, scale, stealth, and ingenuity with which adversaries utilize AI to execute cyberattacks across all phases of the cyber kill chain, as well as the advancement of autonomous offensive tools that the authors refer to in the literature as Recursive Decision Making and Adaptive Evasion as well as large-scale automated attacks. The conclusions of this paper show that traditional means of defence will be increasingly ineffective against future attacks developed and executed through AI by virtue of the inability of defence to detect AI through the gaps and issues associated with current defence systems, including model bias, resource limitations, and lack of transparency in the operation of defence systems. The results of the current study indicate that the shift from human initiated cybercriminal activity to machine initiated and self-learning adversarial systems continues, consequently indicating that an immediate focus must be placed on expanding the proactive and adaptive defence measures needed to deal with a new generation of cybercriminals as well as the enhanced governance and resilience needed to counter the increasing threat posed by these new technologies. Additionally, the present work offers a detailed overview of the current state of knowledge in this domain; identifies the gaps that exist in cybersecurity practices relative to the adoption and use of AI; and discusses how the results of the present study may impact researchers, stakeholders, and the public policy arena as all confront the growing risks posed to both individuals and businesses alike by these emerging technologies.

Keywords - Adversarial AI, AI-generated cybercrime, Autonomous attack platforms, Cyber-attack kill chain, Deep learning-based attacks, Generative Adversarial Networks (GANs), Intelligent malware, Machine learning-enabled malicious attacks, Offensive AI (OAI), Self-directed attacks.

1. Introduction

The use of Artificial Intelligence (AI) in cybersecurity and other fields is gradually but steadily transforming every aspect of our daily lives. While AI helps consumers better protect their networks and data from external threats through improved methods, it also provides cybercriminals with new opportunities to develop more sophisticated attack capabilities that were previously unavailable. Three distinct categories of cybercriminal activity arise from AI's ability to create autonomous attack systems, also known as malware or orchestrated malware, which can learn, plan, adapt, and execute attacks independently with little to no human involvement. The cybersecurityty threat landscape is changing in unprecedented ways and is expected to continue growing rapidly due to ongoing advancements and implementation of AI (Adewale, 2022; Reddem, 2024; Wettstein, 2025). Cybercriminals are now leveraging AI techniques to increase the complexity and frequency of their attacks, develop new tactics to avoid detection, and cause greater damage to their victims (Guembe et al., 2022). Within the evolving cyber threat landscape, AI-generated cyberattacks increased by 238% in 2023 and resulted in total losses exceeding $8.5 billion worldwide.

Agentic AI or self-directed systems are the emerging paradigm of advanced cyber capabilities (Evani, n.d.; Kshetri, 2025) that are used maliciously. These Systems have autonomous characteristics; they can reason independently, set their own goals, and carry out their activities without continuous human supervision (Evani, n.d.; Kshetri, 2025; Wettstein, 2025). The features of these attacks are enabled by AI and include unprecedented speeds and efficiencies through highly advanced and complex CNNs and DNNs (Guembe et al., 2022; Reddem, 2024).

Artificial Intelligence (AI) is inherently dull-minded when it comes to cybersecurity. To put it differently, AI can be seen both as a way to defend ourselves from cybercrime by making it harder for hackers to infiltrate a system and as a new and emerging form of a hacker's attack strategy (Vaid et al., 2023; Walter et al., 2023). Thanks to these technological advancements, there is now a significant difference between the types of threats we face today and those we faced in the past. Our traditional security-based methods for protecting our systems are becoming increasingly ineffective. The speed at which these new types of cyberattacks occur is growing substantially (Guembe et al., 2022; Faheem et al., 2024). Therefore, when hackers leverage developments in AI-enhanced attack techniques, they are able to achieve their goals much faster and with a much higher success rate than ever before (Reddem et al., 2024). Organizations must shift away from their current reactive methods and focus on proactively implementing new, improved AI-enhanced security solutions in order to keep up with and counter these rapidly evolving cyber threats (Guembe et al., 2022; Wettstein et al., 2025).

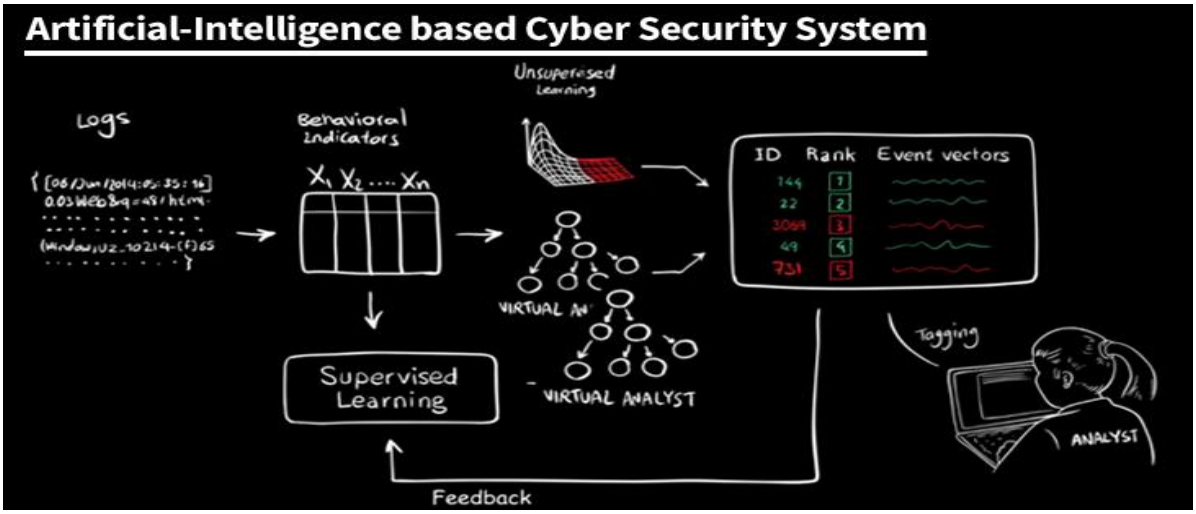

Fig. 1 AI-based cybersecurity system

AI's dual-use capability has accelerated the AI arms race, and defenders must innovate continuously in order to keep up with attackers' growing sophistication and increasing need for AI-based cybersecurity solutions.

Existing research has extensively examined conventional forms of cybercrime, such as phishing, ransomware, and network intrusions, as well as the defensive application of artificial intelligence for anomaly detection, intrusion detection, and threat prediction (Bhatnagar et al., 2018; Kaloudi & Jingyu, 2020). Although prior studies have documented the malicious exploitation of AI, including automated phishing campaigns, deepfake-enabled social engineering, and adversarial attacks targeting machine learning models (Adewale, 2022; Sai Meghana et al., 2024), there remains a notable lack of systematic investigation into fully autonomous and self-directed attack systems.

Much of the existing literature focuses on isolated AI-enabled offensive capabilities, such as vulnerability discovery through machine-learning-based fuzzing or the application of reinforcement learning for lateral movement within compromised networks, without situating these techniques within a comprehensive framework of autonomous cyber operations (Guembe et al., 2022; Kumar & Singh, 2022). Similarly, defensive research has largely emphasized adversarial robustness, anomaly detection, and intrusion detection systems (Ofusori et al., 2024; Rajathi & Rukmani, 2024), while giving comparatively little attention to countermeasures specifically designed to confront adaptive, goal-driven, and self-learning AI-enabled attacks. In parallel, policy-oriented studies highlight the dual-use nature of artificial intelligence and advocate for regulatory oversight; however, they often fallshort of proposing actionable and scalable models for mitigating the risks posed by autonomous and self-direct cyber threats (Brundage et al., 2021).

Consequently, a critical gap exists between advancing offensive AI capabilities, existing defensee mechanisms, and current governance frameworks. Addressing this gap requires a deeper examination of the architecture, functionality, and threat potential of AI-enabled self-directedted attack systems, along with the development of integrated countermeasures that encompass technical, organizational, and policy-level domains.

The goal of this research is to examine the scope and impact of AI-based cybercrime and the development of systems that autonomously coordinate attacks worldwide, including their architecture, how they operatee, and potential defense strategies. To accomplish this, this study will: 1) analyze the evolving threat landscape caused by AI-enabled cybercrime; 2) analyze how self-directed cyber-attack systems are designed and function; 3) explore both actual and hypothetical case studies illustrating various scenarios where AI-enabled attacks might occur; and 4) assess the shortcomings of current defense mechanisms against adaptive and autonomous AI-driven cyber threats. By pursuing these objectives, this research aims to bridge the gap between emerging AI threats and the defense measures needed to counter them.

This study outlines the emerging AI-driven cyber threat landscape, focusing on the shift from human-led cyber-attacks to those automated by machine self-learning systems. Additionally, Section 2 develops a structured methodology for understanding how cybercriminals utilize AI during different phases of a cyber event, including reconnaissance (and exploitation), execution, evasion, and evolution. Furthermore, Section 2 emphasizes the importance of AI-based proactivity through a framework for operational and forensic models to investigate malicious acts involving offensive AI capabilities and the consequences of using such technologies. This comprehensive approach benefits the research community (e.g., academia), policymakers, and the broader cybersecurity field by offering a proactive security strategy.

The review will consist of five separate components. The methodology used to conduct the review will be evaluated in the next segment. Section three will describe the approach taken to carry out the review. Sections four and five will discuss the review’s outcomes and provide closing remarks, respectively.

While prior studies have examined individual applications of artificial intelligence in cybercrime, such as AI-assisted phishing, adaptive malware, and adversarial attacks against machine learning systems (Adewale, 2022; Guembe et al., 2022; Sai Meghana et al., 2024), they essentially treat these capabilities in isolation. Existing reviews primarily focus either on defensive AI techniques or on specific offensive use cases, without conceptualizing how these components converge into fully autonomous, self-directed attack systems capable of independent goal formulationn, adaptive learning, and coordinated execution.

This review is novel in that it systematically synthesizes dispersed research into a unified analytical framework that characterizes AI-enabled cybercrime as an emergent class of autonomous cyber threats. By integrating technical architectures, functional capabilities, defensive limitations, and governance challenges, this work extends the existing literature and provides a holistic perspective currently absent from prior studies.

2. Literature Review

This section discusses the literature related to the study, as well as outlining the research questions and their respective rationale.

2.1. Research Questions and Review Process

This literature review assesses the emergence and impact of AI-enabled cybercrime and self-directed attack systems, while also examining literature relevant to the four Research Questions regarding the review of studies. The rationale for the four Research Questions is located in Table 1 of our literature review, which provided the basis for searching for pertinent studies and creating our criteria for evaluating those studies.

Table 1. Research questions and rationale

Research Question	Rationale
RQ1: What is the current state of AI-enabled cybercrime, and how is it evolving in tactics, techniques, and procedures?	To identify the existing AI techniques used by cyber criminals to cause damage without being noticed in cyberspace
RQ2: What are the key architectural and functional features of self-directed attack systems?	To investigate the architecture and operational mechanism of self-directed attack systems
RQ3 How do AI-enabled attacks exploit autonomy, adaptability, and scalability to bypass traditional cybersecurity defenses?	To evaluate real and hypothetical case scenarios that demonstrate how AI-driven attacks can be deployed in diverse environments
RQ4: What limitations do current defensive approaches have in detecting and mitigating self-directed AI-based cyberattacks?	To evaluate the shortcomings of existing defenses against adaptive and autonomous AI-enabled cyber threats.

2.2. Related Literature

This section presents the related literature to the study thematically by different authors.

2.2.1. The Changing Threat Landscape of AI-Enabled Cybercrime

The growing use of Artificial Intelligence (AI) for malicious purposes has fundamentally changed the global cyber threat landscape, impacting not just the scale and speed of cyberattacks but also their strategic complexity. Recent research consistently shows that AI-enabled cybercrime has led to an exponential increase in potential targets, a significant reduction in attack execution time, and a marked rise in attack complexity, thereby reshaping modern cyber defense strategies (Adewale, 2022; Guembe et al., 2022; Reddem, 2023; Kshetri, 2025). Unlike traditional cyber threats, which generally depend on static scripts or manual orchestration, AI-driven attacks demonstrate adaptive behavior, continuous learning, and rapid optimization, making legacy security measures less effective (Habibu & Julius, 2025)

Beyond technical escalation, AI has introduced substantial economic efficiency into cybercriminal operations. Empirical evidence indicates that AI-enhanced cyberattacks increased by approximately 238% in 2023 alone, contributing to projected global economic losses exceeding USD 10.5 trillion by 2025 (Zandi et al., 2024). Forward-looking assessments indicate a further 400% increase in AI-related attack vectors over an 18-month horizon, with cumulative financial losses projected to exceed USD 1.2 billion by the end of 2025 (Reddem, 2023). These figures underline the systemic nature of AI-enabled cybercrime and highlight the urgency of sustained academic and policy-driven inquiry into its evolving impact.

A key aspect of this transformation is the growing imbalance between offensive and defensive AI applications in cybersecurity. While AI has been widely adopted to enhance threat detection and response, offensive AI systems currently show superior operational efficiency due to fewer ethical, regulatory, and governance constraints (Guembe et al., 2022; Reddem, 2023). Notably, next-generation offensive AI tools have reduced the average Time To Compromise (TTC) from several hours to approximately 19 minutes, representing a 93.1% decrease in execution time (Reddem, 2023). This acceleration has led to a 67% higher success rate for AI-assisted intrusions and a 94% increase in the number of attacks per adversary, increasing both scale and impact.

The rise of agentic and self-directed AI systems represents a further escalation in threat capabilities. Enabled by advances in deep learning, reinforcement learning, and autonomous decision-making architectures, these systems can independently establish objectives, analyze contextual information, and execute coordinated attack sequences with minimal or no human oversight (Kshetri et al., 2023; Klapprohholz et al., 2023). Empirical analyses indicate that the deployment of autonomous attack frameworks can reduce the average time required to compromise a network by up to 85%, fundamentally changing attacker–defender dynamics (Reddem et al., 2023). Such systems blur the distinction between tool-assisted cybercrime and fully autonomous cyber adversaries.

AI has also facilitated the development of advanced Tactics, Techniques, and Procedures (TTPs) across all stages of the cyber kill chain, with particularly pronounced effects during reconnaissance, weaponization, and delivery phases (Guembe et al., 2022; Kazimierczak et al., 2024; Seymour & Tully, 2016). By leveraging machine learning models to analyze historical breach data and behavioral patterns, malicious actors can predict exploitable vulnerabilities and construct highly targeted spear-phishing campaigns optimized through attrition-versus-success analytics. These capabilities significantly enhance social engineering effectiveness while reducing operational costs and human effort.

The rise of Generative AI (GenAI) has further amplified deception-based cybercrime through unprecedented levels of personalization and scalability. Studies report that AI-enhanced phishing campaigns utilizing advanced Natural Language Processing (NLP) achieved success rates of approximately 8.7% in 2023, compared to just 2.9% for conventional phishing methods (Reddem et al., 2023; Zandi et al., 2024). Concurrently, the proliferation of synthetic media, particularly deepfake audio and video, has increased by more than 300%, resulting in reported financial losses exceeding USD 35 million from voice-based fraud incidents alone (Araromi, 2023; Reddem et al., 2023).

Malware ecosystems have also experienced a qualitative shift, driven by AI-enabled adaptability and autonomy. Reports indicate a 43% increase in AI-driven adaptive malware and a 67% rise in autonomously mutating code capable of evading detection (Reddem, 2023). Such malware exhibits continuous learning behavior, environmental awareness, and self-propagation capabilities, enabling it to generate new variants dynamically (Guembe et al., 2022; Kaoudi & Li, 2020). The integration of Deep Neural Networks (DNNs) into malware decision-making processes further complicates detection, as malicious activities increasingly resemble legitimate system behavior (Kirat et al., 2018; Guembe et al., 2022).

Currently, cybercrime has evolved into a mature AI-driven underground economy. The widespread availability of advanced AI tools has significantly lowered the technical barrier to entry, enabling individuals with limited expertise to conduct highly sophisticated attacks (Dixon, 2019; Zandi, 2024). Cybercriminal organizations increasingly resemble conventional technology startups, adopting professionalized structures, specialized roles, and scalable operational models. Underground marketplaces, dark web forums, and encrypted communication platforms now host a wide range of AI-based tools that bypass the ethical safeguards imposed on commercial AI services (Burton, 2025; Klappholz, 2025).

Recent research also highlights the growing prevalence of AI-on-AI attacks, in which cybercriminals exploit vulnerabilities in defensive AI systems. Threat actors use AI-generated TTPs to jailbreak large language models (LLMs), manipulate prompts, and poison training datasets, thereby degrading the performance of AI-based security technologies (Walter et al., 2024; Carlini et al., 2024). Alarmingly, large-scale data poisoning can be carried out at minimal cost, often under USD 60, making it an economically attractive method for undermining AI-driven defenses (Kshetri, 2025).

As AI-driven cybercrime continues to grow, evidence increasingly shows that traditional cybersecurity methods are insufficient. Signature-based detection and reactive defense strategies struggle to keep up with the speed, volume, and novelty of AI-generated threats, especially those using zero-day exploits and polymorphic malware (Guembe et al., 2022; Faheem, 2024; Ofusori et al., 2024). Consequently, experts recommend shifting toward proactive, predictive, and AI-native defense systems capable of anticipating and addressing complex threats in real time (Araromi, 2023; Faheem, 2024; Sarkeer et al., 2020).

2.2.2. How Self-Directed Attack Systems are Built and Operated

Self-directed, or agentic, AI attack systems represent a distinct and increasingly sophisticated class of offensive cyber capabilities characterized by autonomous goal formulation, recursive decision-making, and continuous adaptation. Unlike traditional malware, which executes predefined instructions, these systems rely on complex architectures that enable iterative learning and dynamic behavioral adaptation throughout the attack lifecycle (Evani, n.d.; Kshetri, 2025). The operational effectiveness of self-directed attack systems is therefore determined not by a single algorithm, but by an integrated software architecture that sustains an ongoing cycle of adaptive intelligence.

Architectural Structure and Construction

At the core of self-directed attack systems is the Intelligent Digital Agent (IDA), an autonomous software entity designed to perceive its environment, reason over available information, and execute goal-oriented actions with minimal human supervision (Boston Consulting Group, 2025; Evani, n.d.; Russell & Norvig, 2021). Contemporary IDA architecture is modular by design, allowing for scalability, resilience, and continuous operation across heterogeneous environments. These architectures typically comprise several interdependent functional modules that collectively enable autonomous behavior.

Interface modules establish the communication layer between the IDA and its operational environment, allowing interaction with users, network services, databases, sensors, and external systems. Through these interfaces, the agent can collect environmental data and send commands to target platforms or third-party Application Programming Interfaces (APIs) (Boston Consulting Group, 2025).

Memory modules are critical for maintaining contextual awareness and operational continuity. These modules support both short-term memory, which retains task-specific context during ongoing operations, and long-term memory, which stores accumulated knowledge derived from repeated interactions and prior attack cycles.

Persistent memory allows self-directed systems to refine strategies over time and avoid repeating ineffective actions (Boston Consulting Group, 2025; Evani, n.d.).

The profile module defines the behavioral parameters of the agent, including its objectives, constraints, and operational priorities. This module effectively encodes the attacker's intent by specifying mission goals, acceptable risk threshold, and preferred tactics, thereby guiding autonomous decision-making without continuous human input (Boston Consulting Group, 2025; Evani, n.d.).

Planning modules use artificial intelligence algorithms to create dynamic, multi-stage action plans made up of interdependent subtasks. These plans are constantly adjusted based on environmental feedback, allowing the agent to modify its approach in response to defensive measures or unexpected conditions (Boston Consulting Group, 2025; Evani, n.d.).

Finally, action modules operationalize the generated plans by executing commands, interacting with system APIs, deploying payloads, or coordinating lateral movement within target networks. Together, these modular components form a closed-loop architecture capable of sustained autonomous operation.

Underlying Technical Mechanisms

The intelligence and adaptability of self-directed attack systems are supported by advanced machine learning techniques that embed malicious logic within highly flexible computational models. Deep Neural Networks (DNNs) enable multilayered representation of decision logic, allowing attack behavior to arise from complex internal states rather than explicit rule-based conditions. This architectural opacity makes it increasingly difficultt for defenders to distinguish malicious activity from legitimate system processes, especially when compared to traditional malware based on deterministic "if-then" logic (Guembe et al., 2022; Kirat et al., 2018).

Generative Adversarial Networks (GANs) further enhance stealth by producing adversarial malware variants designed to evade machine-learning-based detection systems. By continuously generating undetectable or low-confidence samples, GAN-enabled attacks can bypass black-box security models and degrade defensive accuracy over time (Anderson et al., 2016; Hu & Tan, 2021; Guembe et al., 2022).

Reinforcement learning (RL) and neural networks are commonly integrated into malware to facilitate adaptive behavior based on continuous interaction with target environments. Through reward-based learning, malicious agents can identify vulnerable services, optimize attack paths, and learn strategies to circumvent authentication mechanisms and access controls (Guembe et al., 2022; Petro & Morris, 2017).

Additionally, fuzzy logic models provide a flexible framework for reasoning under uncertainty, enabling malware to make approximate decisions when precise environmental information is unavailable. This capability supports the developmentt of attack systems that can learn, evolve, and respond effectively to dynamic and ambiguous conditions (Kaloudi & Li, 2020; Guembe et al., 2022).

Operational Processes and Recursive Decision-Making

Self-directed attack systems operate through a continuous feedback loop that supports real-time learning, adaptation, and optimization. Central to this process is a recursive decision-making cycle in which perception, planning, action, and feedback are repeatedly integrated to refine behavior (Evani, n.d.; Kshetri, 2025; Russell & Norvig, 2021).

During the perception phase, the agent continuously monitors its environment using inputs such as system logs, threat intelligence feeds, network telemetry, sensor data, and lateral network exploration. This persistent s

situational awareness enables the agent to detect changes in defensive posture and identify new opportunities for exploitation (Evani, n.d; Russell & Norvig, 2021).

In the goal formulation phase, the agent dynamically prioritizes objectives based on perceived environmental conditions and predefined mission parameters. Goals may be re-ranked or reformulated in response to resistance, detection attempts, or evolving attack opportunities.

The planning and execution phase transforms these goals into multi-step operational strategies, which are carried out through the action modules. Execution is continuously monitored to assess effectiveness and identify deviations from expected outcomes (Evani, n.d.).

Finally, feedback integration ensures that the outcomes of executed actions, both successful and unsuccessful, are reintegrated into the perception phase. This closed-loop learning process enables the system to improve performance across successive attack cycles and adapt to defensive countermeasures (Russell & Norvig, 2021).

Operational Autonomy in Attack Execution

The modular and learning-enabled architecture of self-directed attack systems significantly reduces reliance on human operators, allowing cybercriminals to scale and accelerate attacks with minimal manual intervention (Guembe et al., 2022; Klapprotz, 2025). Self-learning malware can autonomously generate and refine attack strategies, thereby decreasing the level of specialized expertise required by attackers (Chung et al., 2019; Kaloudi & Li, 2020).

Notable examples include Deep Locker-style attacks, in which DNNs conceal malicious payloads and execute only when specific target attributes are detected, therebyy enabling strategic evasion of analysis and sandboxing techniques (Kirat et al., 2018; Guembe et al., 2022). Similarly, AI-enabled Distributed Denial-of-Services (DDoS) attacks can autonomously adjust traffic patterns, target selection, and attack intensity in response to defensive interventions, removing the need for real-time human coordination (Kaloudi & Li, 2020; Reddem, 2024).

In advanced scenarios, AI-enabled malware may operate without traditional Command-and-Control (C2) infrastructures altogether. Instead, autonomous agents can predict optimal payload distribution strategies and coordinate across compromised nodes independently, further complicating detection and disruption efforts (Kirat et al., 2018; Guembe et al., 2022).

2.2.3. Real and Hypothetical Cases Showing How AI-Driven Attacks Can Occur in Different Environments

The literature reviewed identifies numerous hypothetical and actual examples where cybercrime has occurred within various types of physical environments, such as critical infrastructure, Maritime Autonomous Systems (MAS), financial market systems, and the general digital space (e.g., computer systems in the 21st Century). These examples demonstrate the use of AI technology's full capabilities in these environments. The reviewed literature highlights AI's capabilities, including autonomous operation and decision-making (e.g., AI-operated vehicles), the ability to adapt to new and changing operational environments (e.g., AI-enabled financial open markets), and AI's capacity to manipulate human perception and thought patterns through deception or simulation (e.g., simulated entry systems).

Attacks on Physical and Critical Infrastructure Environments

AI-driven attacks pose existential dangers to essential services and critical infrastructures where Cyber-Physical Systems (CPS) govern physical outcomes (Guembe et al., 2022; Walter et al., 2024).

Table 2. Attacks on physical and critical infrastructure

Environment	Attack Type	AI Technique and Mechanism	Impact and Goal	Source
Maritime Autonomous Systems (MAS) (Real-world test case)	Poisoning and Patch-Based Evasion Attacks	Adversarial AI (AAI) utilized a backdoor poisoning attack by injecting data into the training/validation set to associate a visual trigger (a multi-colored flag) with an incorrect classification (Walter et al., 2024). The exploit was delivered via a "rubber ducky/bad USB" to mimic a keyboard operation (Walter et al., 2024).	The goal was to cause the AI anti-collision Dropout Protection Module (DPM) aboard a U.S. Navy experimental vessel (Bauza USV) to misclassify a threat or crash into another vessel when the physical trigger was detected (Walter et al., 2024). This attack bypassed security, appearing normal until the trigger was in the camera's view (Walter et al., 2024).	(Walter et al., 2024)
Cyber-Physical Systems (CPS) (Simulated/Hypothetical)	Intelligent Self-Learning Malware	K-means clustering and Gaussian distribution were utilized to teach malware how to operate and exploit weaknesses without further attacker assistance (Guembe et al., 2022; Chung et al., 2019).	A malicious attack infiltrates and compromises environmental control systems (like cooling systems of a supercomputer facility), masquerading as an unintentional failure on computer infrastructure (Guembe et al., 2022; Chung et al., 2019).	(Guembe et al., 2022)
Cyber-Physical Systems (CPS) (Hypothetical/General)	Autonomous Cyber Attacks	AI-driven systems capable of analyzing network traffic and continuously adapting behavior (Faheem, 2024).	AI-driven attacks could be directed against power grids, water systems, or transportation networks, where compromised digital controls have a direct impact on physical outcomes (Guttieri, 2025; Yohanandhan et al., 2020; Walter et al., 2024).	(Guttieri, 2025; Walter et al., 2024)

Attacks on Financial and Corporate Digital Environments

AI is rapidly evolving the nature of financially motivated cybercrime by automating deception and increasing the efficacy of existing attack vectors (Zandi et al., 2024; Reddem, 2023).

Table 3. Attacks on financial and corporate digital environments

Environment	Attack Type	AI Technique and Mechanism	Impact and Goal	Source
Corporate/Financial Sector (Real-world case)	Deepfake-Based Financial Fraud	Generative AI (synthetic video and audio content) (Burton et al., 2025; Chen & Magramo, 2024).	A worker at a multinational firm in Hong Kong was deceived by a deepfake video and audio impersonating the company's CFO, resulting in a loss of £20 million (Burton et al., 2025; Chen & Magramo, 2024). The attack combined AI deception with conventional phishing attacks (Burton et al., 2025).	(Burton et al., 2025; Zandi et al., 2024)
Financial Sector (Hypothetical/General)	AI-Driven Automated Extortion	Large Language Models (LLMs) are leveraged for automated extortion negotiations (Burton et al., 2025).	Compromising a cloud provider could allow criminal groups to infiltrate hundreds of entities. LLMs could then be used to conduct extortion negotiations on a significant scale, creating a dichotomy of "human-speed defenders versus AI-speed attackers" (Burton et al., 2025).	(Burton et al., 2025)
Cybercrime Marketplaces (Observed/Hypothetical)	Black-Market Crimeware	Specialized LLMs lacking ethical guardrails (e.g., WormGPT, FraudGPT) (Klappholz, 2025; Burton et al., 2025).	These tools, sold on the dark web, streamline criminal processes such as Business Email Compromise (BEC), spear-phishing, cracking, and carding (illicit credit card use) (Burton et al., 2025; Klappholz, 2025).	(Klappholz, 2025; Zandi et al., 2024)

Attacks on General Digital Environments (Web, Social Media, User Systems)

AI enables mass exploitation of software vulnerabilities and human cognitive vulnerabilities across common digital platforms (Guembe et al., 2022; Jameel & Saud, 2022; Reddem, 2023).

Table 4. Attacks on general digital environments

Environment	Attack Type	AI Technique and Mechanism	Impact and Goal	Source
Web/Network Traffic (Real-world data/Simulated)	Automated Phishing Payload Generation	LSTM model (Long Short-Term Memory) (Guembe et al.,	DeepPhish used LSTM to learn patterns from effective phishing	(Guembe et al., 2022)

		2022; Bahnsen et al., 2018).	URLs and generate new synthetic phishing URLs that evade AI detection systems (Guembe et al., 2022; Bahnsen et al., 2018). This raised attack success rates from 0.69% to 20.9%, and from 4.91% to 36.28% in trials (Guembe et al., 2022).	
User Authentication Systems (Simulated/Hypothetical)	AI-Driven Password Cracking/Guessing	Generative Adversarial Networks (GANs) (Guembe et al., 2022; Hitaj et al., 2019).	PassGAN evolved an automated technique by learning password distribution from actual breaches (Guembe et al., 2022; Hitaj et al., 2019). This method correctly predicted between 51% and 73% more unique passwords than traditional brute-force methods (Guembe et al., 2022).	(Guembe et al., 2022)
Targeted Social Media Users (Simulated/Hypothetical)	Automated Spear-Phishing/Disinformation	Markov chains, LSTMs, and NLP (Natural Language Processing) (Guembe et al., 2022; Seymour & Tully, 2016).	Automated systems crafted personalized, machine-generated disinformation posts on Twitter aimed at high-value users, adapting content based on the target's posting history to ensure maximum persuasive potential (Guembe et al., 2022; Brundage et al., 2018; Seymour & Tully, 2016).	(Guembe et al., 2022; Seymour & Tully, 2016)
AI Systems/LLMs (Observed/Hypothetical)	Jailbreaking and Guardrail Removal	Prompt manipulation	Attackers manipulate LLMs to	(Burton et al.,

		techniques (Burton et al., 2025; Yigit et al., 2024).	bypass ethical guardrails to elicit information about criminal strategies, including how to develop malware or assist in crime planning (Burton et al., 2025; Yigit et al., 2024). Criminals are developing their own LLMs with guardrails removed (Burton et al., 2025).	2025; Yigit et al., 2024)
Recruitment Systems (Hypothetical)	AI Bias and Discrimination	Biased AI algorithms trained on unrepresentative data (Slimi, 2023).	AI systems used in recruitment can produce discriminatory hiring outcomes by reproducing ethnic or gender bias when screening candidates (Slimi, 2023).	(Slimi, 2023)

In conclusion, these cases illustrate that AI-driven cybercrime is using machine efficiency for scale that is automating phishing and cracking, adaptability that is malware self-learning and goal inversion, and deception that is deepfakes and sophisticated social engineering across virtually every digital and physical domain (Guembe et al., 2022; Reddem, 2023).

2.3. Comparative Analysis

The following systematic review table summarizes the key architectural and functional features, applications, findings, and limitations of influential studies about AI-enabled cybercrime and countermeasures, as cited and discussed throughout the literature.

Table 5. Systematic literature review summary

Author and Year	Focus of the Study	AI Techniques Used	Key Findings	Gap/Limitation	Relevance to the AI-enabled Cybercrime
Guembe et al. (2022)	Systematic review of the emerging threat and techniques utilized in AI-driven cyberattacks (Offensive AI).	Deep Learning, GAN, CNN, RNN, LSTM, K-means clustering, Fuzzy Model.	The study revealed that 56% of AI-driven cyberattack techniques were demonstrated in the access and penetration	Organizations need to invest in AI cybersecurity infrastructures to combat emerging threats.	Direct analysis of offensive AI Techniques, Tactics, and Procedures (TTPs) across the Cyber Kill Chain.

			phase. Existing cyber defense infrastructures are inadequate due to the increasing speed and complex decision logic of AI-driven attacks.		
Reddem (2023)	Data-driven analysis of the surge, impact, and efficiency of AI-powered cyber attacks.	AI, Machine Learning (ML), Generative AI (implied by deepfake/phishing focus).	The study shows that AI-driven attacks saw a 238% surge in 2023, achieving a 67% higher success rate and reducing the average time-to-compromise from 4.6 hours to 19 minutes.	Traditional defense strategies are inadequate and require immediate, comprehensive evolution to counter AI's efficiency.	Quantifies the massive acceleration and high success rate of AI in automating criminal operations and amplifying financial impacts.
Kshetri (2025)	Examination of the transformative potential and security risks of agentic AI (autonomous agents) in cybersecurity.	Agentic AI, Large Language Models (LLMs), Adaptive Algorithms.	The study reveals that Agentic AI systems can autonomously detect and respond to threats; conversely, malicious agentic AI could autonomously strategize, reason, and execute multi-stage operations.	Current autonomous attack agents remain unreliable, requiring significant refinement for large-scale criminal deployment.	Defines the emerging threat of fully autonomous, self-directed attack systems capable of deep decision-making.
Kazimierczak et al. (2024)	Systematic review of adversarial AI strategies and their impact	AI, GANs, Deep Learning.	AI-based tools are most effective in the initial stages of cyberattacks	Limited research compiling both offensive and defensive techniques	Provides a structured understanding of where AI provides the greatest

	on each stage of the Cyber Kill Chain (CKC).		(Reconnaissance, Weaponization, Delivery, and C&C). Current defense tools are inadequate for countering sophisticated AI attacks in these early stages.	across all stages of the CKC.	advantage to attackers within the offensive lifecycle.
Walter et al. (2024)	Proposed RED-AI framework for evaluating the security of AI in Maritime Autonomous Systems (MAS) against adversarial AI (AAI).	Adversarial AI (Poisoning, Evasion, Backdoors), CNN (YOLOv8 model).	The study shows that AAI is a "present-day threat"; vulnerabilities were found across the AI development lifecycle. Successfully demonstrated a backdoor poisoning attack against a real-world MAS anti-collision model.	Attack and defense methods evaluated in low-entropy labs differ in complex, dynamic real-world environments.	Direct assessment of AAI capabilities targeting mission-critical Cyber-Physical Systems (CPS), demonstrating risks of physical harm.
Shinde et al. (2024)	Systematic review of Blockchain solutions for securing AI-based healthcare systems against adversarial attacks.	NLP, Computer Vision (CV), Acoustic AI (Application targets); DNN; GAN (Attack technique).	AI models face challenges from insufficient data, adversarial attacks (poisoning and evasion), and a lack of trust due to their opaque black-box nature. Adversarial attacks pose a severe risk of misdiagnosis.	Existing defense techniques for adversarial attacks are often specific to one attack vector and are themselves AI-based, making them susceptible to adversarial exploits.	Details specific mechanisms of adversarial attacks that corrupt data integrity and compromise predictive AI models in critical services.
Anderson, Woodbridge	Proposed GAN-based	Generative Adversarial	GAN-generated undetectable	Focused on specific black-	Demonstrated a highly effective AI

, & Filar (2016)	automatic generation of undetectable malware URLs (AI-Concealment Attack).	Network (GAN).	malware URLs successfully bypassed DNN-based malware detection systems and traditional classifiers.	box malware detection systems.	TTP for malware evasion/concealment in the Delivery and Command & Control (C2) phases.
Hu & Tan (2021)	Proposed generating adversarial malware examples for black-box attacks (Intelligent Concealment).	GAN.	Developed a GAN technique capable of generating undetectable adversarial malware to bypass machine learning black-box detection systems.	Focused on black-box detection systems.	Specific AI TTP for malware evasion, maximizing stealth against defensive ML models.
Lee & Yim (2020)	Cybersecurity threats based on machine learning offensive techniques for password authentication via keystroke analysis.	LR, SVM, SVC, RF, KNN, GBRT, MP (Machine Learning Classifiers).	The model accurately predicted and stole actual user passwords based on keyboard strokes with 96.2% accuracy.	Limited research on AI defenses designed to generate realistic dummy keystrokes.	Illustrates AI use in Reconnaissance/Access for high-accuracy information theft by exploiting human-machine interaction patterns.
Ofusori et al. (2024)	Comprehensive systematic review of AI techniques and applications in cybersecurity.	Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP).	AI is indispensable for anomaly detection, threat identification, and automated response. This reflects a trend of accelerating complexity in defense mechanisms.	There is a lack of in-depth studies comparing which AI models are most effective for specific cybersecurity tasks under real-world conditions.	Provides the foundational context for defensive AI adoption and highlights the escalating complexity of the cybersecurity arms race.
Rahman et al. (2024)	Evaluation of Linear (LC) vs. Nonlinear Classifiers (NLC) for	LC (LR, NB, LDA, etc.) and NLC (DT, RF, GBM, XGB, etc.).	NLC significantly outperformed LC (up to 23.68%	The developed scheme lacks feature extraction using Deep Learning	Demonstrates that AI-driven cyber threats are structurally complex, validating

	Intrusion Detection Systems (IDS) robustness.		accuracy increase in unseen data) because intrusion detection data exhibits nonlinear characteristics.	models.	the necessity for advanced NLC models to counter them.
Rajathi & Panjanathan (2025)	Development of a two-phase feature selection framework (2P-FSID) to improve IDS performance and interpretability.	Feature Selection, SHAP values (Explainable AI), Logistic Regression.	The framework reduced dimensionality (e.g., 41 to 19 features) while maintaining high accuracy, increasing computational efficiency, and significantly enhancing explainability.	Need for further validation on larger, real-world datasets and broader exploration of other XAI approaches (LIME, etc.).	Highlights the growing importance of interpretability and transparency (XAI) in building trust in complex cyber defense models, essential for human oversight.

2.3.1. The Shortcomings of Current Defense Mechanisms against Adaptive and Autonomous AI-Driven Cyber Threats

Recent literature consistently concludes that the contemporary cybersecurity threat landscape has undergone a structural transformation driven by advances in adaptive, autonomous, and AI-enabled attack mechanisms. Across technical, organizational, and strategic studies, researchers identify a widening asymmetry between the speed, adaptability, and autonomy of offensive AI systems and the comparatively rigid, human-dependent nature of prevailing defensive frameworks (Guembe et al., 2022; Reddem, 2023; Vaid, 2023). Rather than attributing this imbalance to a single technological shortcoming, the literature frames it as a multidimensional failure encompassing obsolete detection paradigms, exploitable weaknesses in defensive AI itself, and persistent operational constraints that inhibit effective deployment at scale (Faheem, 2024; Kshetri, 2025; Wettstein, 2025). Collectively, these studies suggest that current cybersecurity architectures were largely designed for deterministic, human-orchestrated adversaries and therefore struggle to contend with threats characterized by machine-speed decision-making, continuous learning, and autonomous goal pursuit. The following sections synthesize the literature across these dimensions to demonstrate the depth and scope of the identified gap.

The Obsolescence of Traditional, Rule-Based Defenses

A substantial body of literature documents the declining effectiveness of traditional, rule-based, and signature-driven security mechanisms in the face of AI-enabled attacks. Early intrusion detection and prevention systems were optimized for identifying known attack patterns and anomalies derived from historical data. However, as Guembe et al. (2022) and Reddem (2023) emphasize, such systems exhibit structural rigidity that renders them ineffective against adaptive and previously unseen threats. Empirical studies quantitatively highlight this limitation. Reddem (2023) reports that signature-based IDS platforms achieve approximately 76% detection accuracy for known threats, yet performance degrades sharply—to roughly 23%—when faced with zero-day exploits. This decline is particularly pronounced in scenarios involving polymorphic malware, which autonomously mutates its codebase to evade static signatures and heuristic rules (Guembe et al., 2022). These findings are reinforced across multiple studies that characterize legacy defenses as inherently reactive, relying on

post-compromise indicators rather than proactive threat anticipation. The literature further emphasizes the time mismatch between attackers and defenders. AI-driven automation has significantly shortened the attack lifecycle, reducing the average time-to-compromise from several hours to just minutes (Reddem, 2023). Burton et al. (2025) conceptualized this imbalance as a contest between "AI-speed attackers" and "human-speed defenders," noting that even well-resourced security operations centers cannot consistently respond within such tight timeframes. As a result, defenses relying on human-in-the-loop validation or post-event forensic analysis are becoming increasingly outdated. Beyond exploitation, researchers also highlight weaknesses in early-stage detection. KaziMierczak et al. (2024) and Yamin et al. (2021) demonstrate that adversaries now use AI during reconnaissance and delivery phases of the Cyber Kill Chain, employing predictive analytics and automated probing to identify high-value vulnerabilities with minimal footprint. Existing defense tools, which mainly focus on payload inspection or anomaly detection during execution, are often blind to these non-invasive, preparatory activities. This allows attackers to optimize their strategies long before triggering traditional alarms.

Vulnerabilities Inherent in Defensive AI Systems

In response to the limitations of legacy defenses, organizations increasingly deploy AI-driven security solutions. However, a growing strand of literature warns that these defensive AI systems introduce a new class of vulnerabilities that sophisticated adversaries actively exploit. This phenomenon, commonly referred to as Adversarial AI, has emerged as a central concern in recent cybersecurity research (Kshetri, 2025; Shinde et al., 2024; Walter et al., 2024). Multiple studies demonstrate that machine learning models, especially deep neural networks, are highly vulnerable to adversarial examples. Small, often unnoticed manipulations of input data can cause a model to misclassify malicious activity as harmless, effectively bypassing detection without changing the underlying attack logic (Guembe et al., 2022; Shinde et al., 2024). This vulnerability undermines the assumption that AI-based defenses inherently enhance robustness. The literature also highlights risks associated with the training and lifecycle management of defensive models.

Baker (2025) and Kshetri (2025) document how data poisoning attacks allow adversaries to corrupt training datasets, causing models to internalize flawed decision boundaries. Backdoor or Trojan attacks are particularly insidious: by embedding hidden triggers during training, attackers can ensure the model behaves normally under standard conditions while failing catastrophically when specific patterns are present (Walter et al., 2024). These attacks are difficult to detect through conventional validation techniques, further complicating trust in defensive AI. Emerging research extends these concerns to agentic and autonomous AI systems. As security platforms adopt autonomous agents capable of independent decision-making, scholars warn of goal inversion and objective manipulation attacks. Evani (n.d.) and Kshetri (2025) describe scenarios in which malicious inputs alter an agent's reward or oobjective function, redirecting its behavior in ways that actively undermine defensive goals. This represents a qualitative shift in risk, as compromised agents may operate at scale and speed without immediate human oversight.

Operational and Technical Barriers to Integration

Beyond technical vulnerabilities, the literature identifies systemic operational challenges that limit the real-world effectiveness of AI-enabled defenses. Ofusori et al. (2024) and Vaid (2023) argue that the complexity of advanced AI models introduces governance, trust, and resource allocation issues that are often underexplored in purely technical studies. One recurring theme is the lack of explainability in modern AI systems. Deep learning models often function as "black boxes," providing limited insight into how specific security decisions are reached (Guembe et al., 2022; Jameel & Saud, 2022). This opacity hinders human oversight, making it difficult to determine whether anomalous behavior reflects adversarial compromise, data drift, or normal variance. As Wettstein (2025) notes, this uncertainty erodes organizational trust and limits the willingness of practitioners to grant autonomous systems greater control. The problem is caused by alert fatigue. Several studies report that AI-driven detection systems generate excessive false positives due to noisy or irrelevant features in training data (Tiwari et al., 2020;

Rajathi & Panjanathan, 2025). Over time, security analysts become desensitized to alerts, increasing the likelihood that genuine threats will be overlooked—a phenomenon consistently observed in operational environments (Vaid, 2023). Finally, the literature highlights a persistent "defensive lag" between adversarial innovation and defensive maturity. Adewale (2022) and Guttiepu (2025) note that criminal exploitation of AI usually outpaces the development, testing, and standardization of corresponding defenses. This gap worsens due to methodological limitations: many defensive approaches are validated in controlled, low-entropy laboratory settings that fail to capture the complexity and volatility of real-world networks (Sommer & Paxon, 2010; Ofusori et al., 2024). High implementation costs and a global shortage of cybersecurity professionals—estimated at nearly four million—further restrict the scalability of advanced defensive solutions (Kshetri, 2025; Tambe et al., 2020).

2.4. Synthesis and Identified Gap

Taken together, the reviewed literature shows that the inadequacy of current cybersecurity defenses is not due to isolated technical flaws but to a systemic misalignment between rapidly evolving, autonomous attack capabilities and defensive architectures constrained by legacy paradigms, exploitable AI models, and operational bottlenecks. While individual studies suggest incremental improvements, there remains a notable absence of holistic frameworks that integrate resilience, explainability, and autonomy without increasing risk. This unresolved gap underpins this review.

2.5. Contribution beyond Existing Literature

This review advances beyond recent literature by integrating offensive and defensive AI research into a unified, adversarial systems perspective rather than treating them as separate domains. It challenges the prevailing assumption that AI-based defenses are a straightforward improvement, foregrounding adversarial AI vulnerabilities as core design constraints rather than peripheral risks. Additionally, it extends existing work by incorporating operational and governance limitations, such as explainability, alert fatigue, and skills scarcity, into the main analytical framework.

3. Materials and Methods

This section describes how the systematic review of the literature was conducted by the authors.

3.1. Systematic Review Methodology

To ensure rigor, transparency, and reproducibility, this study employed a Systematic Literature Review(SLR) methodology grounded in established evidence-based research guidelines in computer science and cybersecurity. The review was designed to comprehensively identify, evaluate, and synthesize peer-reviewed research on the capabilities of AI-enabled cyber threats and self-directing attack systems, as well as the effectiveness and limitations of the current defensive framework.

3.2. Review Design and Rationale

A systematic review approach was chosen over a traditional narrative review to reduce selection bias, ensure broad coverage of relevant literature, and enable clear justification of inclusion decisions. Given the rapidly evolving nature of AI-driven cybersecurity, the review highlights recent, high-impact studies while still incorporating foundational works where necessary to contextualize emerging trends. This studys systematic literature review method was based on PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), a guideline designed to improve the reporting of systematic reviews (PRISMA, 2020). The research aimed to identifyy relevant studies on AI-enabled cybercrime and self-directed attack systems. To achieve this, the authors outlined the study's research questions and explained the rationale for each. They also discussed the criteria for selecting releva nt literature and the search method used to retrieve pertinent publications.

3.3. Search Criteria and Identification of Studies

The following search parameters were used to find relevant literature for this study:

- Make a list of keywords from the research questions.
- Identify keywords in relevant literature.
- Recognize distinct keyword synonyms and spellings.
- To relate primary keywords and concepts using the Boolean operators "AND" and "OR."

The search keywords are developed from the research questions in Table 1. The output of the search string used for searching relevant literature is as follows: ("AI-driven cyberattacks" OR "AI-driven attack techniques" OR "malicious use of AI and self-attacking systems") AND ("AI-powered cyberattacks" OR "Self-attacking systems" OR "Artificial intelligence in cyberattack" OR "Impact of AI-Driven attack"). Three factors were used to apply the eligibility criteria: inclusion, exclusivity, and quality criteria. These criteria were used to extract the literature from the search results.

3.3.1. Exclusive Criteria

The following exclusive criteria were used to evaluate the retrieved literature:

- EC1: Non-discussion of the research questions in the literature.
- EC2: Articles on the same subject.
- EC3: The same articles from different databases.
- EC4:Articles that do not discuss AI-driven cyberattacks.

3.3.2. Inclusion Criteria

The retrieved literature was evaluated with the following inclusive criteria:

- IC1: Literature that is relevant to AI-driven cyberattacks.
- IC2: Methodologies, Journals, Conferences, and White Papers that addressed AI-driven attacks.
- IC3: Papers address AI-driven attack techniques such as deep learning, bio-inspired swarm intelligence, etc.

3.3.3. Quality Criteria

The selected papers were screened based on the following quality criteria:

Quality Criteria

The selected papers were evaluated against the following quality criteria:

- QC1: To what extent did the paper(s) address most research questions?
- QC2: Is there a detailed description of the AI-driven attack type(s) and the corresponding AI-based technique(s) to execute this (these) type(s) of attack?
- QC3: Did the study answer most of the research questions?
- QC4: Did the paper(s) adequately describe the methodology(ies) used by researchers in the studies?

3.4. Data Sources

The primary source of the data obtained from this research was the following databases and collections: ACM, arXiv, Blackhat, Scopus, Springer, MDPI, and Taylor & Francis Online.

3.5. Selection Procedure

The primary search sources for this study included ACM, arXiv, Blackhat, Scopus, Springer, MDPI, Taylor & Francis Online, and IEEE research databases. The PRISMA flowchart (Figure 2) is used to illustrate the systematic review process and the subsequent selection of relevant papers at different stages of the selection. In the AI-enabled Cybercrime and Self-Aware Systems field, the first stage of the initial systematic review process (Information Extraction) involved conducting a rigorous and systematic search through the eight identified databases (the eight

electronic databases). The outcome of this stage was 936 identified article outlines that served as a source of possible articles, as in Table 6. In the second stage, the screening stage, a total of 936 were identified, as in Table 6. There were 417 duplicates identified because many articles were identified across multiple online databases in Stage 1.

During the screening process, the titles of the articles were reviewed for apparent relevance. At this stage, 309 articles were determined to be inappropriate for use in the study. In the third stage, eligibility was evaluated by confirming the relevance and quality of the articles. An article's relevance cannot be determined just by looking at the title or abstract; therefore, all articles evaluated in this systematic review used complete information extraction criteria based on complete text reviews. At this point in the study, 210 articles were eligible for consideration, and 164 articles were eliminated due to vague, non-specific techniques. In the fourth stage, the author's inclusion was based on a quality assessment of the research on the above topics for the remaining papers. After resolving issues with article selection, a total of 46 primary papers were reviewed for this investigation.

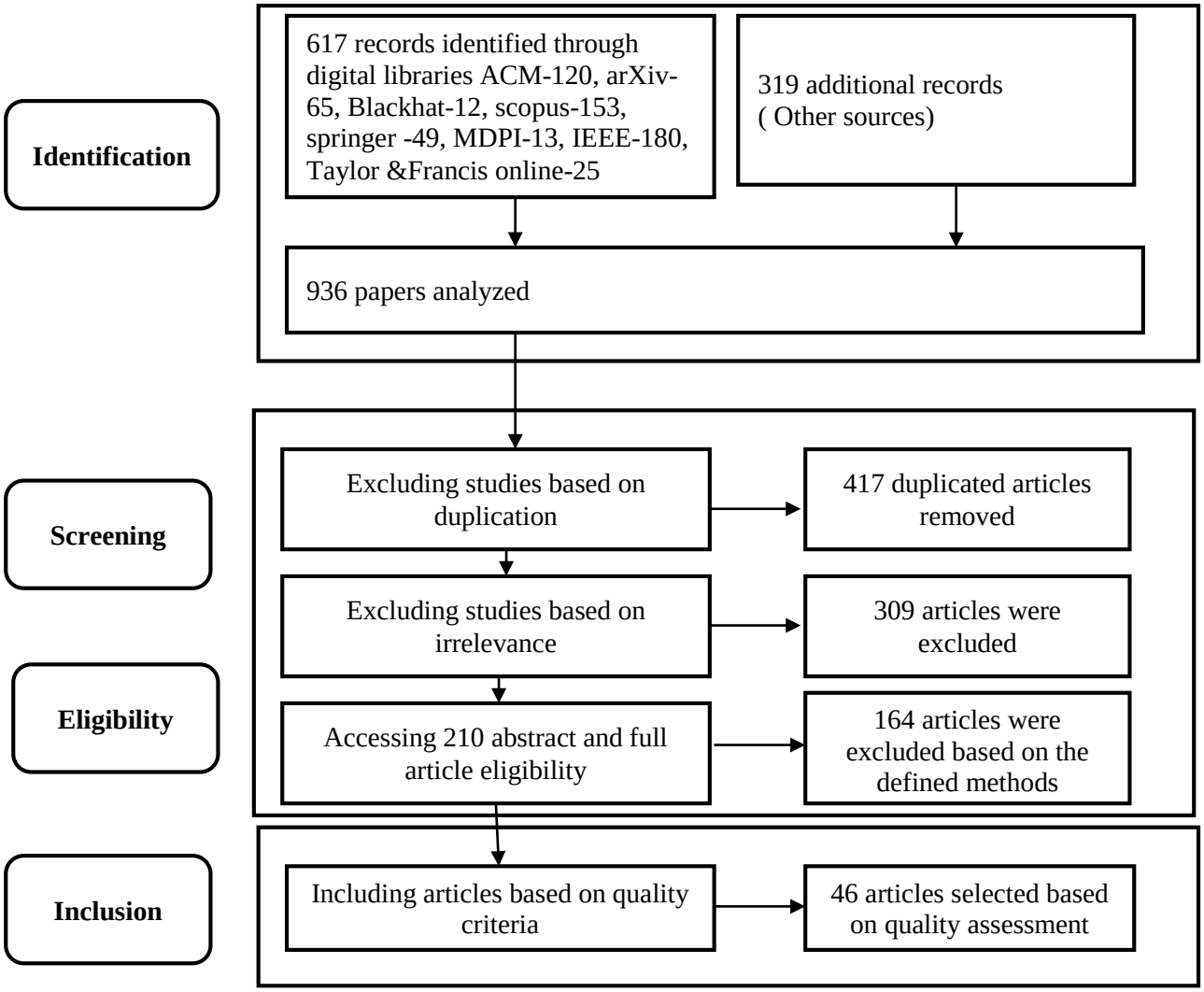

Fig. 2 PRISMA flowchart showing the systematic review process and article selection at different stages

Table 6. Number of Publications Identified in Online Databases

S/N	Database	No. of articles
1	ACM	120
2	arXiv	65
3	Blackhat	12
4	Taylor & Francis online	25
5	Scopus	153
6	Springer	49
7	MDPI	13
8	IEEE	180
9	Others	319

3.6. Search Strategy

This section describes and analyzes the different search strategies used to select the 46 papers for this study. The distribution of the articles is as follows: fifteen (15) papers from journals, fourteen (14) from conference proceedings, eight (8) from workshops, six (6) from symposiums, two papers (2) from white papers, and one (01) from scholarly work, as shown in Figure 3. This study categorized the search approach into four groups to examine all contributions from previous researchers in this field. The four search techniques used are techniques, status, source, and attack strategy.

3.6.1. Sources

In this study, data were collected from eight digital databases. ACM, arXiv, Blackhat, Scopus, Springer, MDPI, Taylor & Francis Online, and IEEE Xplore are accessible digital libraries. Conference proceedings, journal papers, workshops, symposiums, and scholarly works were searched using titles, abstracts, and keywords.

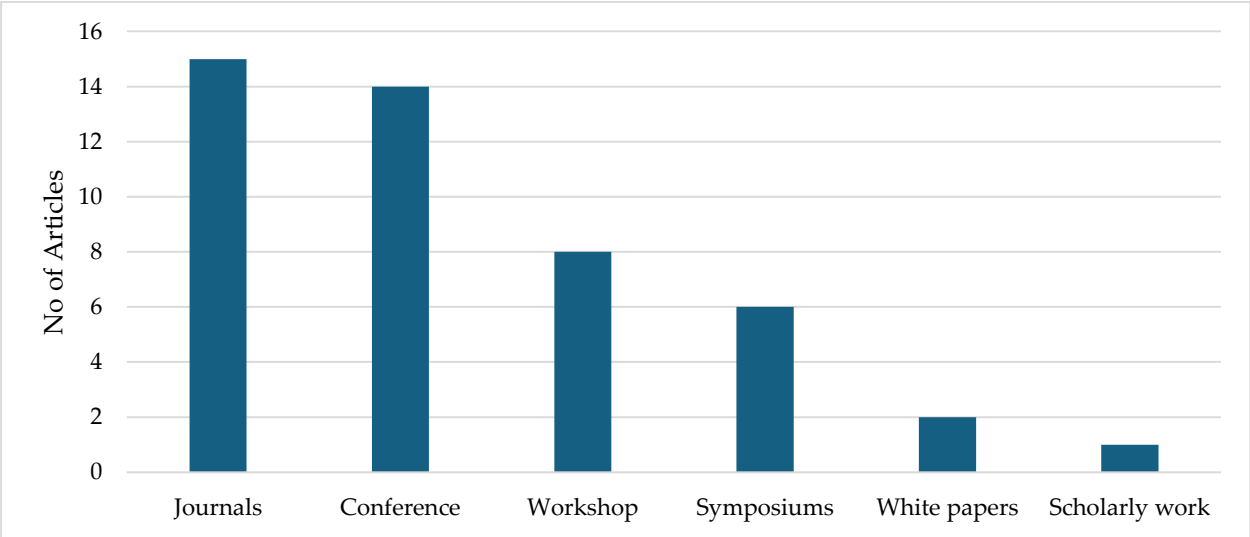

Fig. 3 Number of collected studies

3.6.2. Attack Strategy

In this paper, the AI-driven attack strategies identified in the forty-six selected papers include deep learning, bio-inspired computation, swarm intelligence, and fuzzy models. Many of the selected papers focus on the deep learning strategy as described in this section.

Types of Attacks

Based on the 46 selected papers, this section identifies nineteen use cases of offensive AI across six stages of the cybersecurity kill chain, as shown in Figure 4. In the access and penetration phase (AI-aided attack), six types of AI-driven attacks are identified. Four types occur in the access reconnaissance stage (AI-targeted attacks), three in the exploitation stage (AI-automated attacks), two in the delivery stage (AI-concealment attacks), and two in the C2/Action on Objectives stage (AI multi-layered attacks). In contrast, one type of AI-driven attack appears in the Action on Objectives stage (AI malware attack), as shown in Figure 4. Figure 5 shows that the access and penetration stage has the most publications (6), followed by the reconnaissance stage (4). The exploitation stage has three publications, and the delivery and C2 stages each have two. In contrast, the action-on-objectives stage has the fewest publications (1).

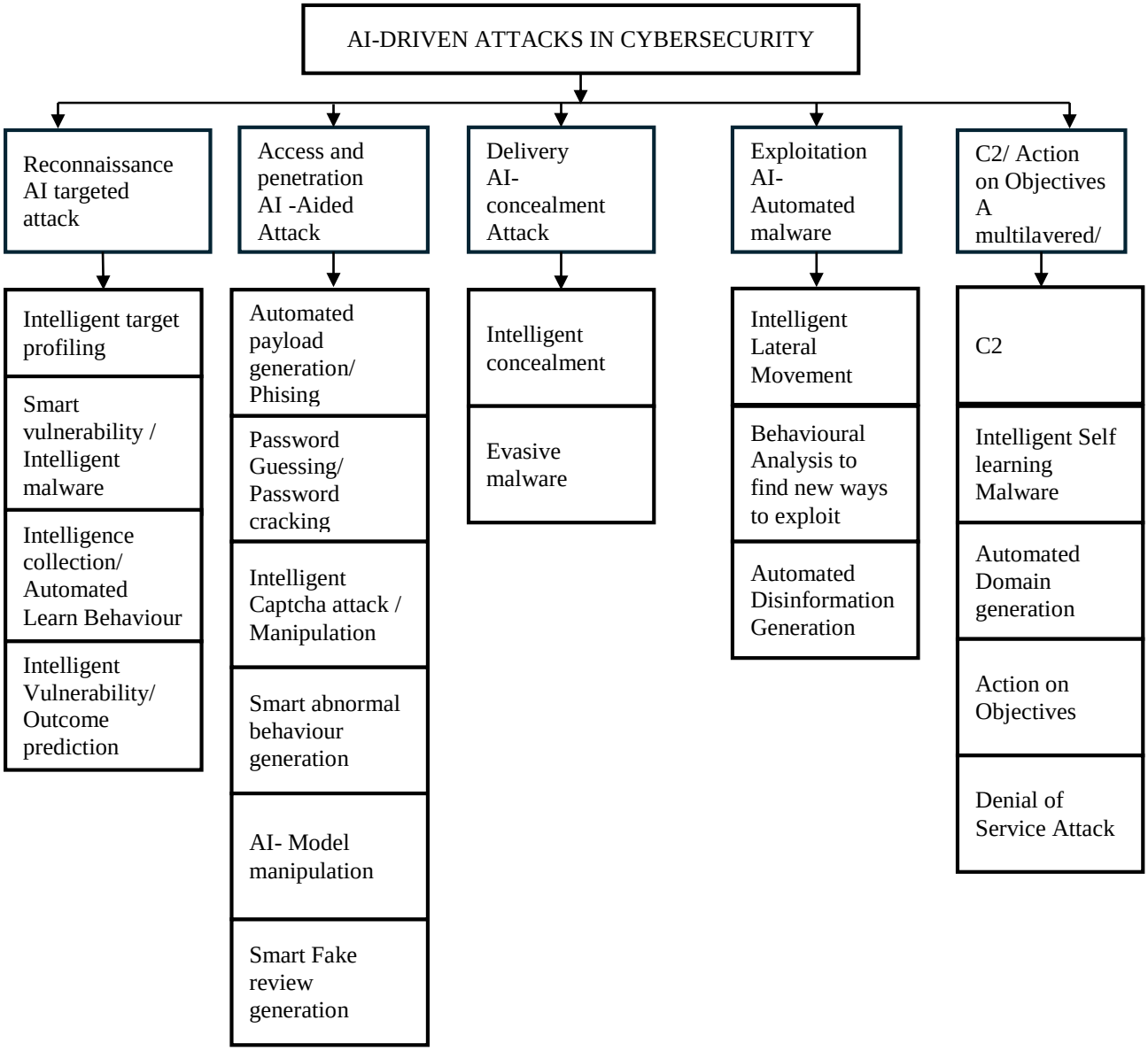

Fig. 4 Modified cybersecurity kill chain for AI-driven attacks (Kaloudi and Li, 2020)

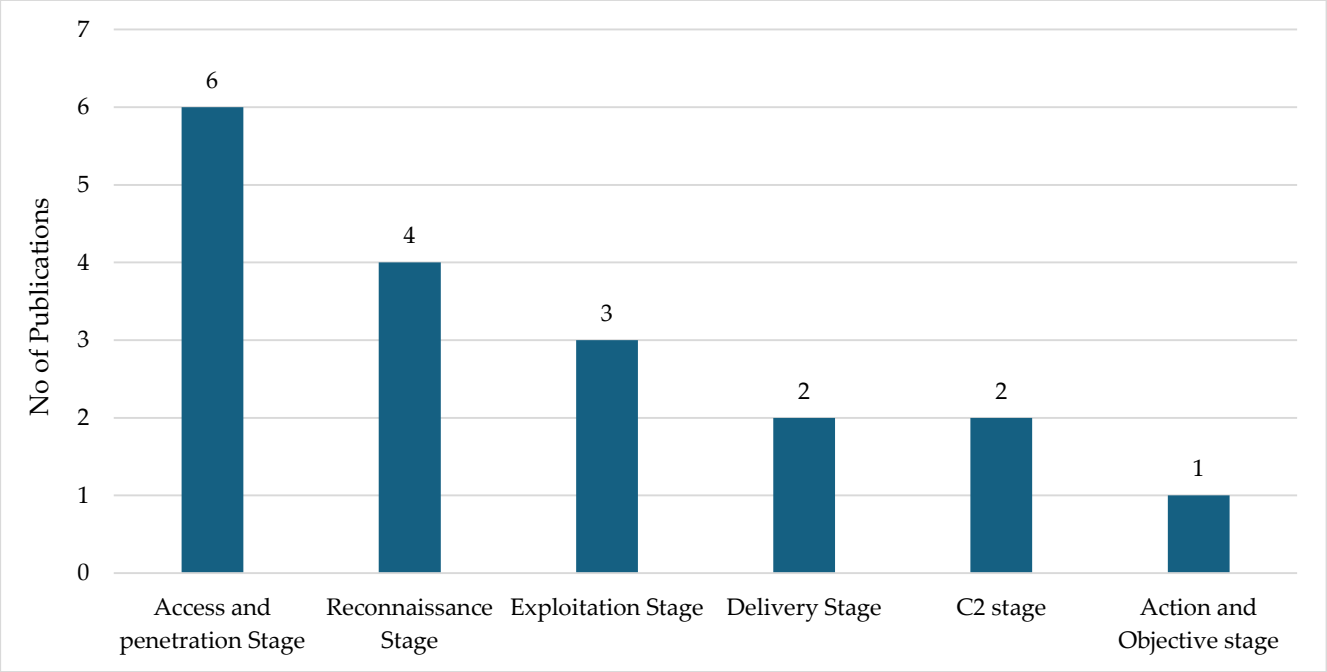

Fig. 5 Offensive AI techniques used in cyberattacks in a modified cybersecurity kill chain

3.6.3. Techniques

According to selected studies, malicious actors can employ AI techniques in their attack process. Using AI-based approaches, malicious operators predicted the multitude of vulnerabilities, exploited those weaknesses, and profiled their victims' online presence by analyzing data and performing reconnaissance (or the "intelligence gathering" phase) of the cyber kill chain. Many studies listed the AI-based techniques used by the authors under the two categories of access and penetration in their own studies. CNN was found to be the most common technique used by the authors with respect to the access and penetration attacks. Only a few authors employed GAN and RNN to perform these types of attacks. All other AI techniques, including LSTM, SVC, SVM, cycle GAN, TOD+ CNN, RF, MP, GBRT, and KNN, had one or more who performed an attack depiction utilizing that technique. Three of the articles listed AI techniques that were utilized in the delivery phase of the kill chain.

GAN was used to conduct intelligent concealment and generate adversarial software (malware) and undetectable software (malware) URLs as well as LSTMs to generate automatic, evasive, and malicious payloads; these methods are stated in separate studies. Finally, one study mentions that DNNs can hide malicious intent and later activate those malicious intentions when victimizing specific individuals (Kirat, Jang, and Stoecklin, 2018). The references that were chosen to discuss 3 (three) types of AI that demonstrated how behavior could be analyzed to uncover new ways to breach target systems and create automated misinformation about their targets. In the first article, Neural Networks (NNs) and Reinforcement Learning (RL) are used to assess the behavior of users in order to determine their weaknesses when using the web application, ultimately enabling someone to bypass authentication controls.

The selected papers also demonstrated how K-Means Clustering could show that AI-assisted self-learning malware can effectively take advantage of security detection systems' weak points by mimicking unintentional failures to attack and take over sensitive environmental control systems as part of the exploitation step of the Cyber Kill Chain. Markov chains and LSTMs were used to generate automated, machine-produced disinformation via an end-to-end spear-phishing technique, creating personalized content for high-value targets on Facebook. Two types of AI-driven cyber-attacks were identified: intelligent self-learning malware and automated domain generation.

Four AI techniques were identified as tools malicious actors might use to execute AI-driven self-learning malware and automated domain-generation attacks during the command-and-control phase of the cybersecurity kill chain. Two of the studied approaches utilized K-means clustering, Gaussian distribution, and DNNs to demonstrate intelligent self-learning malware attacks. Overall, 56% of the AI-driven cyberattack techniques were observed during the access and penetration phases; 12% were identified in the exploitation and command-and-control phases; 11% in the reconnaissance phase; 9% in the delivery phase; and no AI techniques were shown to be active during the action on the objective stage of the cybersecurity kill chain, as illustrated in Figure 6.

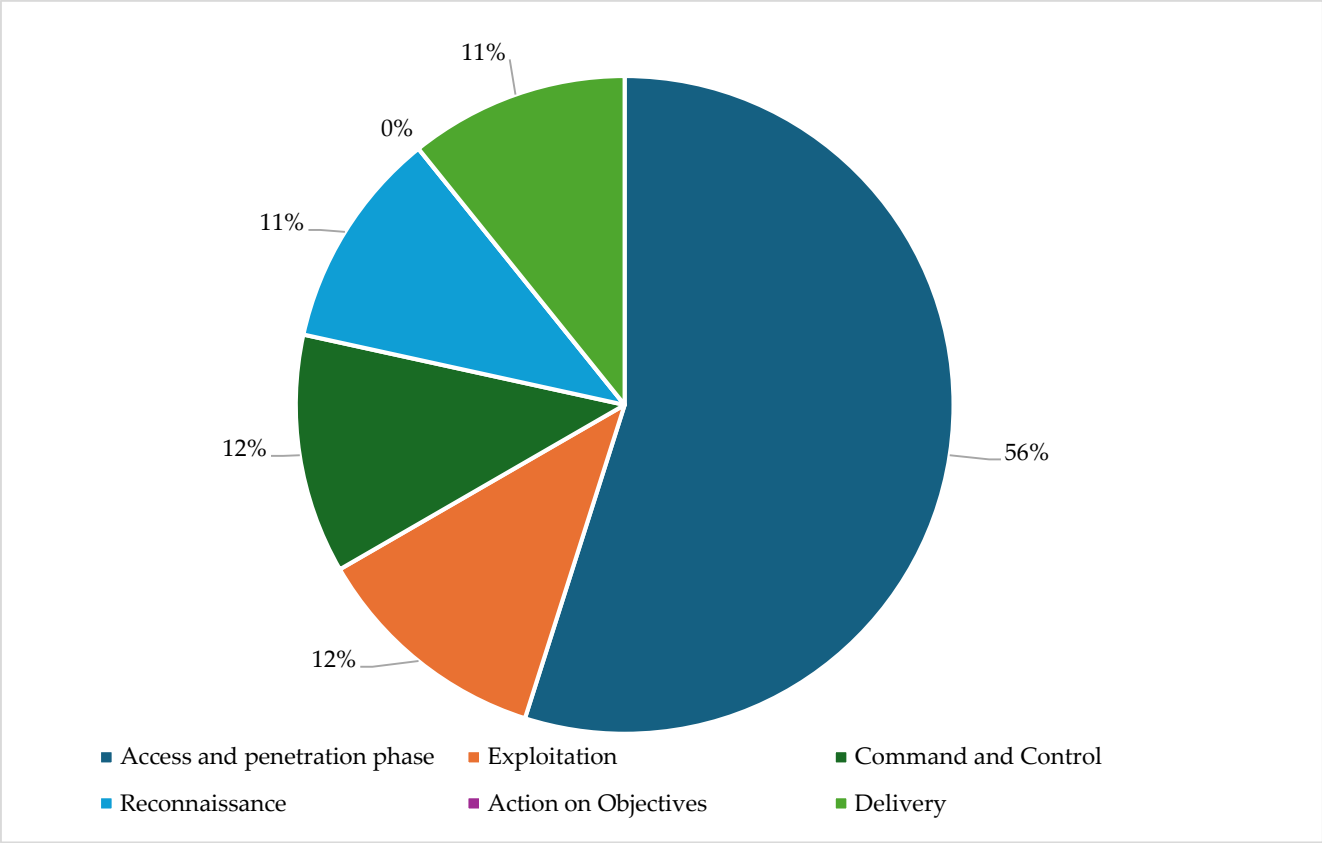

Fig. 6 Detected AI-powered cyberattack techniques

3.6.4. Status

The literature chosen is grouped into three categories. This is illustrated in Table 7.

Table 7. Overview of existing literature and its respective description

S/N	State	Description
1.	Implemented	This refers to a category of studies that designed a proof of concept to demonstrate AI-driven attacks.
2.	Proposed	This category of studies is based on new techniques or methods without any proof of concept or evaluation.
3.	Implemented and evaluated	This category of studies refers to those that designed a proof-of-concept demonstrating AI-driven cyberattacks and evaluated it using performance metrics.

4. Results and Discussion

This section presents the study's findings and provides a comprehensive discussion.

4.1. Results Obtained

This section analyzes the results of the four search methods in detail and discusses the study's findings. Several subsections provide a comprehensive discussion of the results in relation to the study's topics, along with concise interpretations. A word cloud analysis of the titles of selected articles shows 'Learning' as the most common term, followed by 'Artificial', 'Machine', 'attack', 'attacks', 'self-directed attacks', and 'cybersecurity.'

4.1.1. Search Strategy 1: Source

The initial search exercise used a hierarchical approach to identify related articles on AI-driven cyberattacks and self-directed attack systems, using article titles and keywords before developing a final search strategy. The following databases were used to locate relevant literature for publications published between 2020 and 2025: ACM, arXiv, Blackhat, MDPI, Scopus, Springer, Taylor & Francis Online, and IEEE Xplore. The findings for relevant article sources are shown in Figure 7, while the number of relevant publications released during the research year is shown in Figure 8.

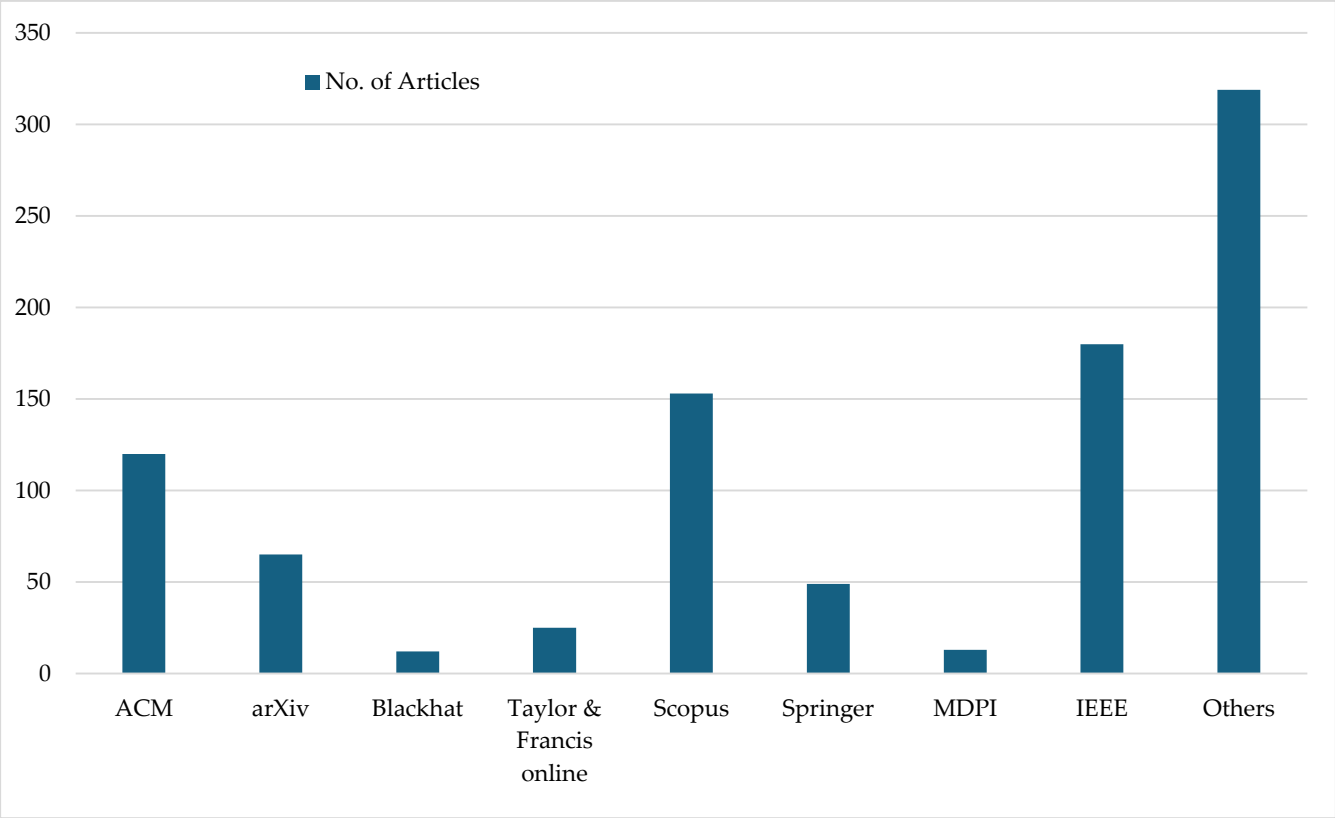

Fig. 7 Average number of relevant articles

Figure 8 shows that out of the final 46 papers, ACM had the most relevant publications with eleven (11), followed by arXiv with five (5). Blackhat and IEEE each had four (4) publications, while Scopus and Springer each had two (2). Taylor and Francis Online contributed three (3) publications, and the remaining fifteen (15) relevant publications were retrieved from other sources.

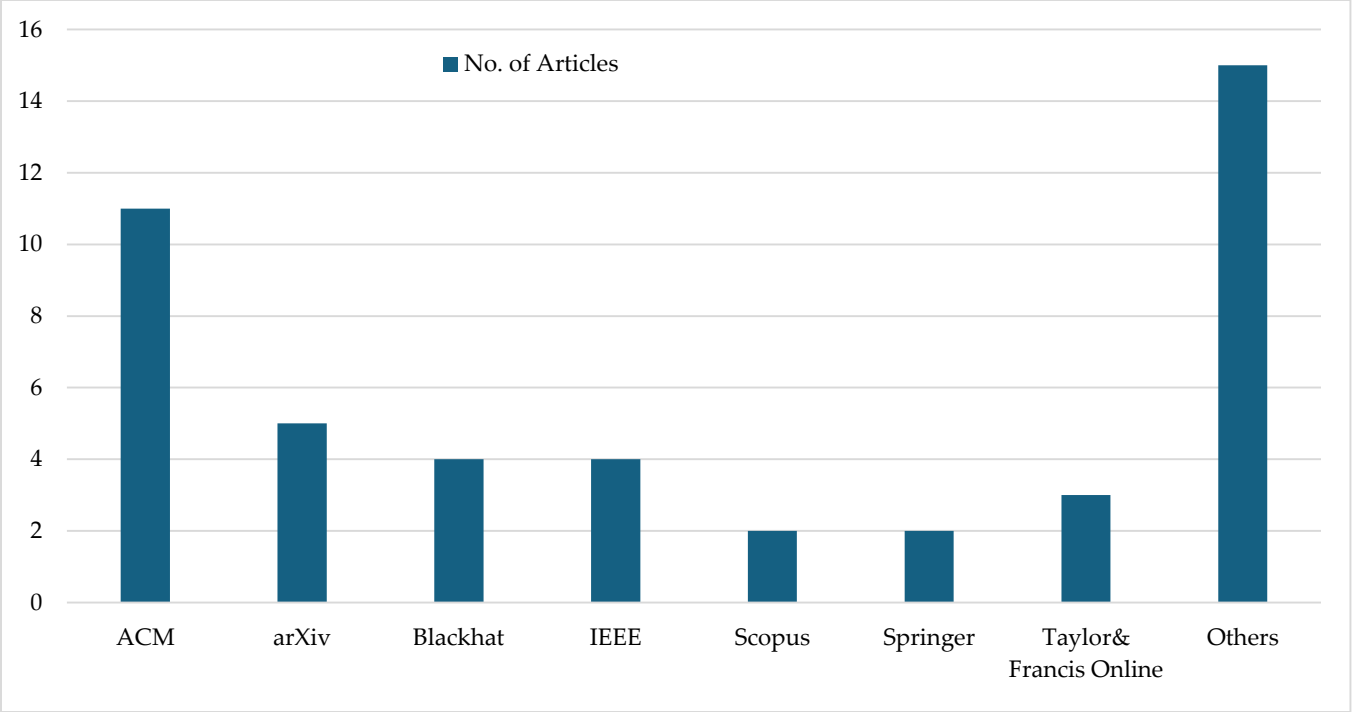

Fig. 8 Relevant articles by database

Figure 9 shows that 2024 had the most relevant papers (14), while 2020 had the fewest publications (2).

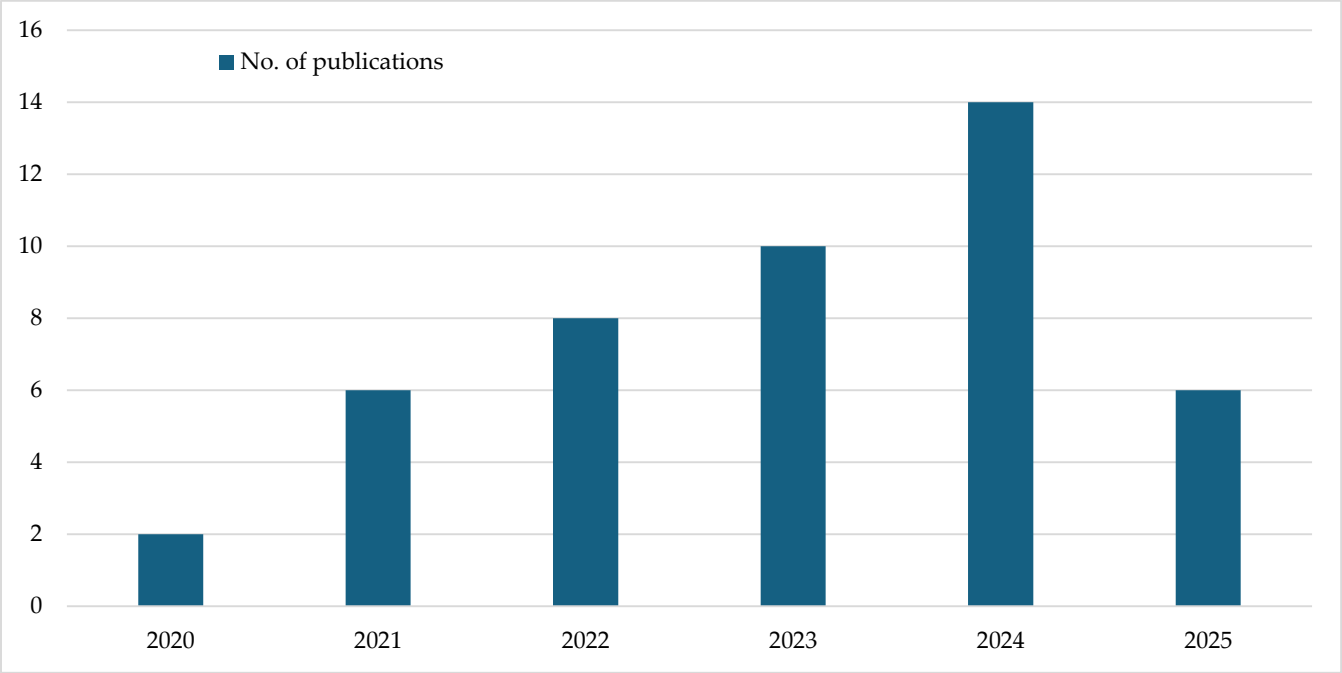

Fig. 9 Year of publication

The threat of AI-driven cyberattacks increases despite ongoing research to understand and counter these advanced cyber weapons. Figure 10 illustrates the AI techniques used by the selected studies to show the malicious application of AI in cyberattacks during the access and penetration phase of the modified cybersecurity kill chain.

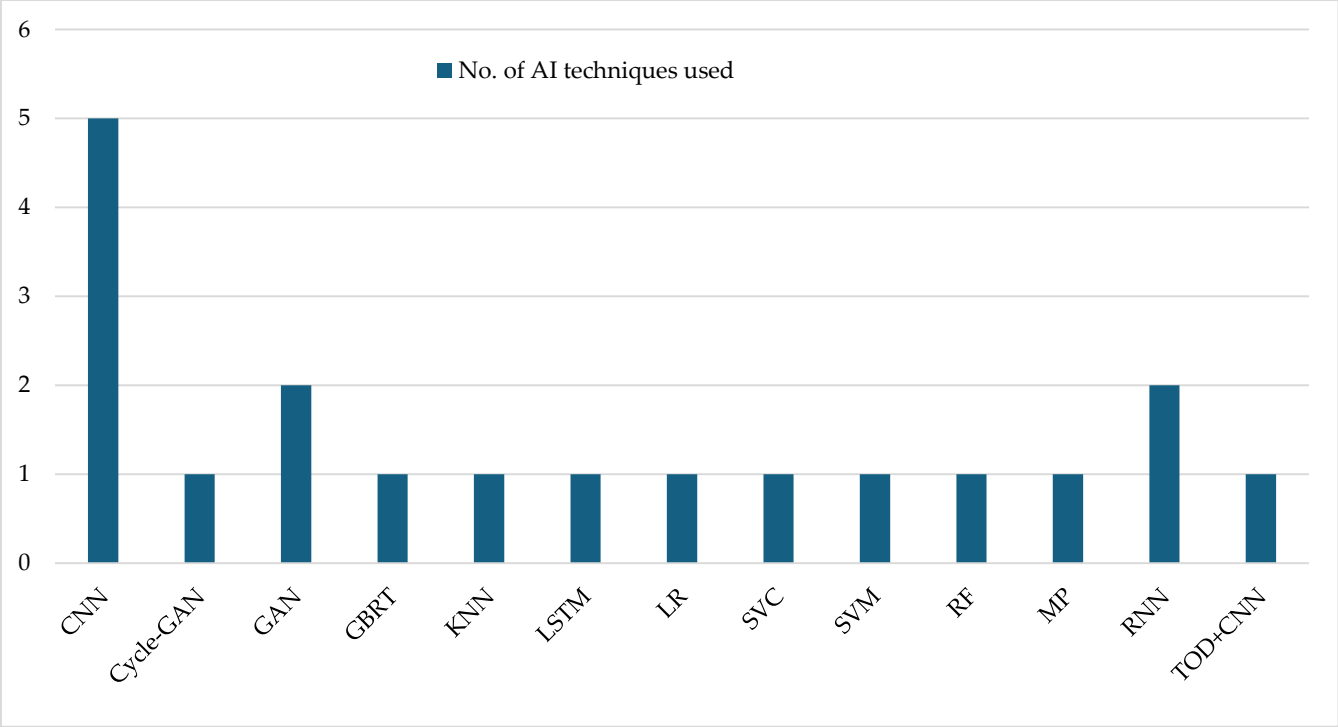

Fig. 10 AI Methods in the access and penetration attack phase

Figure 11 depicts the AI techniques used by the selected authors to demonstrate the malicious use of AI in the delivery stage of the modified cybersecurity kill chain. The results indicate that GAN has the most publications (2), while DNNs and LSTM have one 1 publication each.

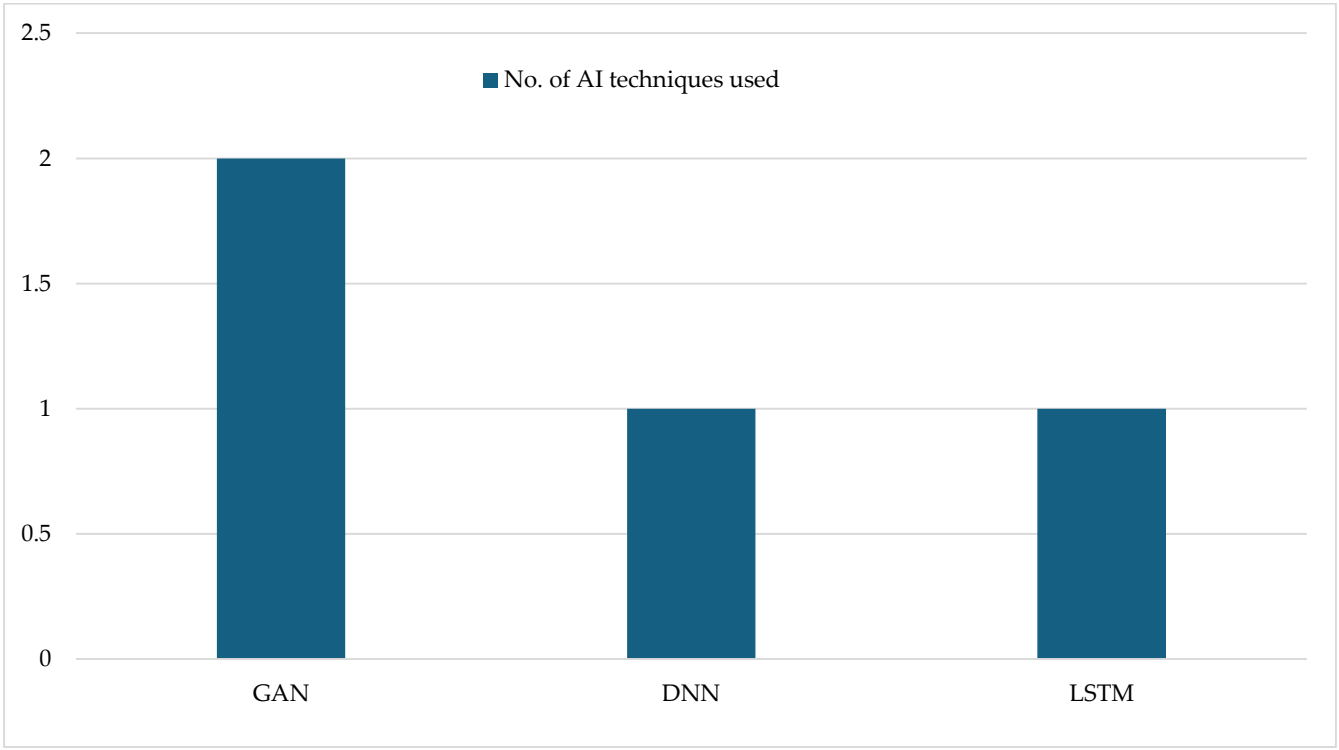

Fig. 11 AI techniques in the Delivery stage.

4.1.2. Search strategy 2: Techniques

This section's findings are organized into six stages of the cyber security kill chain: reconnaissance (AI-targeted attack), access and penetration (AI-aided attack), delivery (AI-concealment attack), exploitation (AI-automated malware), command and control (AI-multi-layered attack), and action on objectives (AI-massive attack).

Reconnaissance Stage

During the reconnaissance Stage, three AI techniques were found within the Cyber Kill Chain. Selected studies showed how cybercriminals utilized Markov chains/LTSM, NNs, and DNNs to perform: Vulnerability Prediction, E2E Spear Phishing, and Intelligent Target Profiling/Intelligence Collection.

Access and Penetration Phase

A total of six (6) AI-Based Attacks have been identified by this research that takes place in the access and penetration stages. The six (6) types of AI-Based Attacks include: Password Guessing/Cracking (Brute Force), Intelligent CAPTCHA Manipulation, Intelligent Generation of Abnormal Behavior, AI Manipulation of Models, and Creation of Fake Reviews.

In addition to identifying six (6) types of AI-based attacks in the Access and Penetration Stages, this research has also identified 19 different AI Techniques that will be used by Malicious Actors to perpetrate Access/ Penetration Attacks. As illustrated in Table 5 of this research, AI techniques will be used by Malicious Actors when conducting these types of attacks on the access and penetration stage.

Delivery Stage

Two forms of intelligence-based cyberattacks against companies were disclosed that fall within the Delivery Phase of the Cyber Kill Chain Model. These two forms include either Evade A.I. or Intelligent Concealment. In addition, based on the studies we reviewed, the following are examples of how Cybercriminals can use intelligent concealment and evasive malware through the Delivery Phase of the Cyber Kill Chain Model.

Exploitation Stage

The exploitation phase involves gaining authorized access to computer applications and resources. After access is gained to the target, malicious actors can use AI techniques to carry out complex attacks that are hard to detect with NNs and DNNs.

Command and Control Stage

In the Command and Control (C2) phase, two types of AI-driven cyberattacks were identified: intelligent self-learning malware and automated domain generation.

4.1.3. Search Strategy 3: Status

The current state of the AI-driven attack tool was assessed using three categories, and the results of the analysis for the selected publications are shown in Figure 12. Based on existing AI-driven cyberattack technologies, the analysis results for the 46 articles evaluated in this study indicate that 63% of the AI-driven cyberattack tools are implemented and evaluated, 25% are proposed only, and 12% are implemented without evaluation. Overall, most existing AI-driven cyberattack tools are based on implementation and evaluation.

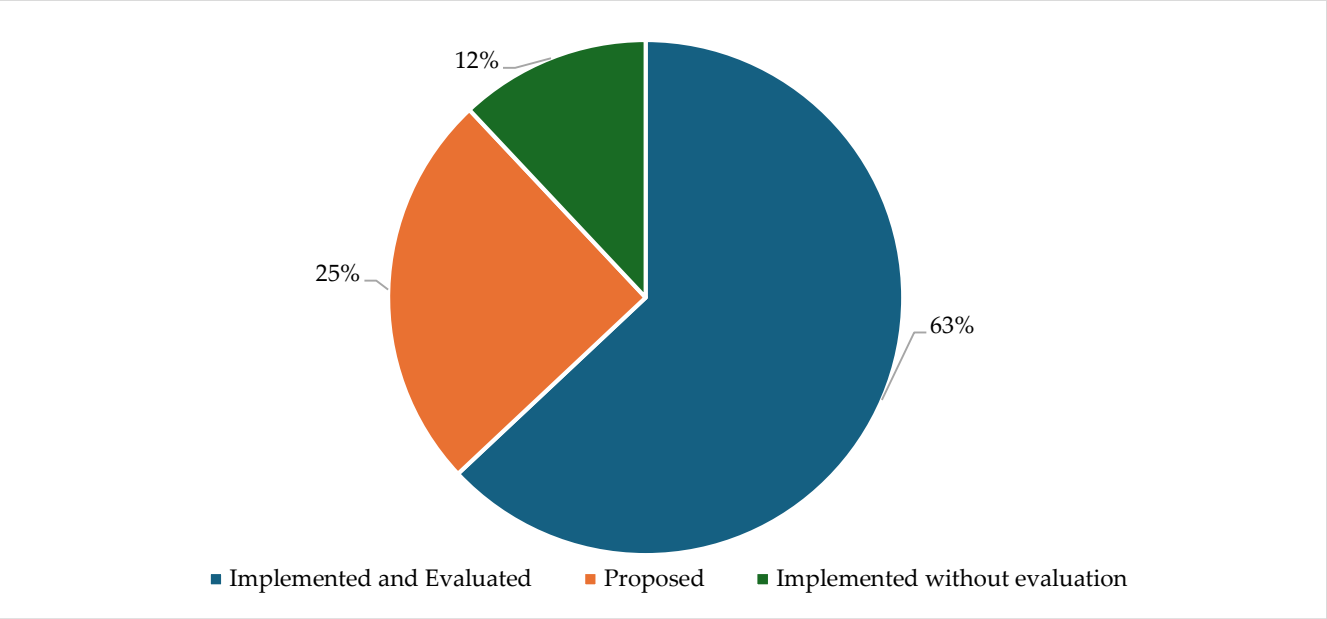

Fig. 12 Status of existing AI-driven cyberattack tools

4.2. Discussion on Research Questions

This section reviews the core concepts that underpin this study, including current and emerging AI-driven cyberattack techniques, the operational and architectural characteristics of self-directed attacking systems, types of AI-driven attacks in the cybersecurity kill chain, and existing AI-driven attacks, how AI-enabled attacks bypass traditional cybersecurity defenses, and limitations in the existing defensive approaches for detecting and mitigating self-directed AI-based cyberattacks.

4.2.1. RQ1: What is the Current State of AI-Enabled Cybercrime, and How is it Evolving in Terms of Tactics, Techniques, and Procedures?

The development of cyberattack tools and techniques is shaping and expanding the cyberattack landscape, exposing cyberspace to a wide range of sophisticated cyber weapons with significant adverse effects (Kaloudi & Li, 2020). Guembe et al. (2022), in their systematic review investigating emerging threats and techniques of AI-powered cyberattacks (offensive AI), found that the main techniques used included Deep Learning (CNN, GAN, RNN, LSTM, DNN), Neural Networks (NNs), Reinforcement Learning (RL), K-means clustering, Markov chains, Gaussian distribution, bio-inspired swarm intelligence, and fuzzy models. The findings of Guembe et al. (2022)'s study revealed that 56% of the identified AI-driven cyberattack techniques were used during the access and penetration phase of the Cyber Kill Chain (CKC), which is an AI-assisted attack.

Furthermore, AI-enabled methods can change or adapt according to their surroundings and take advantage of flaws in systems. The methods used in their research that made use of AI were CNN techniques that appeared in access/penetration attacks a total of 5 times each. Additionally, the methods developed based on AI for access and penetration attacks have proven to be incompletely effective compared to traditional defense structures due to the speed and complexity with which AI methods are able to operate (Guembe et al., 2022; Yamin et al., 2021).

In addition, the status of cybercrime using AI is characterized by rapid advancement in sophistication, scale, and speed of attack, forcing a transformation of existing cyberattack TTPs to allow for more automation and adaptability than ever before (Adewale, 2022; Reddem, 2024). The integration of AI into malicious activities is an exceptional change in the digital threat arena, allowing for a more compensatory attack against traditional security measures (Guembe et al., 2022; George, 2024; Vaid, 2023). AI can both be used and challenged as a "double-edged

sword" by those committing crimes (Adewale, 2022; Reddem, 2024). AI–enabled cybercrime has accelerated growth and financial impact, thus AI-driven cyberattacks rose by 238% in 2023, with global damage expected to exceed 10.5 trillion annually by 2025 (Morgan, 2020, as cited in Prince et al., 2024; Zandi et al., 2024).

AI-enabled cyber-attacks have improved efficiency and speed, and as a result, AI-powered attacks have a 67% higher success rate than traditional methods while reducing operational complexity for attackers by 72% (Reddem, 2024). AI-driven automation has significantly shortened the average time-to-compromise from 4.6 hours to just 19 minutes (Reddem, 2024).

The growing trend of AI-enabled cyberattacks is worsening the cyber threat landscape. AI-driven cyberattacks are expected only to increase, making them nearly impossible for traditional cybersecurity tools to detect, mainly because these tools cannot keep up with speed, complex decision logic, and diverse variants of AI-driven attacks (Guembe et al., 2022; Kazimierczak et al., 2024).

AI is being maliciously applied across all stages of the Cyber Kill Chain, augmenting existing crime types and lowering barriers to new entrants (Araromi, 2024; Burton et al., 2025; Guembe et al., 2022).

Reconnaissance and Targeting (Planning Phase)

AI techniques are used to improve efficiency during the initial planning and intelligence-gathering stages (Kazimierczak et al., 2024; Guembe et al., 2022). Malicious actors use AI, including Deep Neural Networks (DNNs) and Neural Networks (NNs), to identify and profile targets (Guembe et al., 2022; Yamin et al., 2021)

Weaponization and Concealment

Focusing on AI's role in the automation, stealthiness, and highly deceptive implementation of malware (Guembe et al., 2022; Kazimierczak et al., 2024) in terms of the creation of highly convincing phishing emails and BECs via Generative AI (GenAI) and Large Language Models (LLMs) by removing grammatical errors and mimicking natural language (Zandi et al., 2024). The use of AI has increased the success rate of phishing attacks to 8.7%, compared to the previous 2.9% (Reddem, 2023; Zandi et al., 2024). GenAI can generate synthetic media, such as deepfakes, which are realistic forgeries created using audio or video (Zandi et al., 2024; Yigit et al., 2024). The recent growth of deepfake technology has led to a 300% increase in the number of recorded incidents associated with cybercrime in 2023, resulting in total losses associated with voice fraud of approximately $35 million (Reddem, 2023).

New Generation Cyber Threats are more intelligent and can operate autonomously with the aid of AI (Guembe et al., 2022) due to the incorporation of AI-based Adaptive mechanisms into malware evolution by 43%, and AI-based Autonomous Code Modification by 67% for evasion of detection (Reddem, 2023; Guembe et al., 2022). An example of evasive malware is Deep Locker, which conceals its payload with Deep Neural Networks (DNNs) and activates only when specific target characteristics are detected (Guembe et al., 2022; Kirat et al., 2018).

Delivery and Access/Penetration

The Access and Penetration phase accounts for the largest share of AI-driven attack techniques identified in some research (56%) (Guembe et al., 2022). Generative AI (GenAI) and Large Language Models (LLMs) are revolutionizing the development of personalized, sophisticated social engineering campaigns (Zandi et al., 2024; Adewale, 2022). AI models generate phishing messages and Business Email Compromise (BEC) communications that mimic natural language and eliminate grammatical errors, making them much more challenging to detect (Zandi et al., 2024). An AI algorithm utilizing Long Short-Term Memory (LSTM) models learns patterns from effective phishing URLs to generate new, synthetic, undetectable phishing URLs, significantly increasing success rates (Bahnsen et al., 2018, as cited in Guembe et al., 2022).

Specialized illegal tools like WormGPT and FraudGPT, available on the dark web, are specifically designed to help craft convincing BEC communications and improve spear-phishing attacks (Burton et al., 2025; Zurier, 2023, as cited in Burton et al., 2025). GenAI is used to produce highly realistic synthetic media, including video and audio deepfakes, to impersonate individuals such as executives, enabling sophisticated scams (Adewale, 2022; Reddem, 2024; Zandi et al., 2024). This multimodal approach, combining synthetic video and audio, was demonstrated in a notable 2024 Hong Kong financial fraud case involving a deepfake of a company's Chief Financial Officer (Burton et al., 2025; Chen & Magramo, 2024, as cited in Burton et al., 2025).

Neural networks, especially GANs, have revolutionized password cracking (Reddem, 2024). PassGAN uses deep learning to automate password guessing by analyzing real password breaches (Hitaj et al., 2019, as cited in Guembe et al., 2022). AI-powered attacks can try 2.7 million passwords per second, far surpassing traditional methods (Reddem, 2024).

Exploitation and Command and Control (C2)

The exploitation phase of the AI-Automated Malware relies on AI for persistent access, self-learning, and masquerading as benign activity (Guembe et al., 2022). AI-driven malware is built using fuzzy models or K-means clustering to monitor the target system, self-launching operations by strategically injecting an attack at the most vulnerable time and location (Guembe et al., 2022; Chung et al., 2019). This self-learning capability reduces the knowledge needed by the human attacker (Guembe et al., 2022). Intelligent malware can infiltrate and compromise cyber-physical systems (CPS), such as environmental control systems, while masquerading as unintentional failures on computing infrastructure to avoid detection (Guembe et al., 2022; Chung et al., 2019).

AI enables attackers to create powerful C2 channels or eliminates the need for them by using autonomous systems. AI is being used by attackers to create pseudo-randomly generated malicious domains using GANs to evade DL-based detectors and other ageing technologies, thus enabling stealthy C2 communications (Guembe et al., 2022).

Attackers use GANs to generate a large volume of malicious domains that successfully evade detection by Deep Learning (DL)-based detectors and other older forms of machine learning systems (Anderson et al., 2016). With the increasing use of Artificial Intelligence (AI) technologies (Guembe et al., 2022), specifically in malware creation, the possibility of independently predicting when and how to execute its payload eliminates conventional C2 channels altogether. AI systems can make autonomous decisions regarding the timing and methods used to conduct attacks, based on their knowledge of the target's current attributes (Kirat et al., 2018).

Action on Objectives

Due to the processes converging, objectives will be completed at machine speed and at the same scale (Guembe et al., 2022; Kazimierczak et al., 2024). In turn, this removes the need for any human input; therefore, as the defense reacts to an attack against it, malicious AI systems can immediately change their methods of vulnerability exploitation and how they attack the defense (Guembe et al., 2022; Kaloudi & Li, 2020). As a result of sophisticated financial fraud becoming increasingly prevalent, AI tools can conduct negotiations on extortion at scale, whether it has been through compromising a cloud service provider or through using stolen data (Burton et al., 2025).

4.2.2. RQ2: What are the Defining Architectural and Functional Characteristics of Self-Directed Attack Systems?

Architecture

The defining architectural and functional characteristics of systems that are able to direct attacks on their own are also known as "self-directed attack systems" or "agentic Artificial Intelligence" by some. The architecture of a self-directed attack system consists of several intelligent digital agents that contain intelligent modules that perform specific functions, which means that they can independently engage in complex activities without requiring the

continual intervention of a human (Evani, n.d; Kshetri 2023). The Self-Directed Attack Systems are designed from an architectural perspective as digital agents, with each digital agent having an array of various types of functionally distinct and independent modules, enabling these systems to autonomously and continuously perform the process of self-direction. Self-Directed Attack Systems comprise several interconnected modules that have been designed, from a systems perspective, to enable the autonomous operation of the agent. As such, Self-Directed Attack Systems contain:

The self-directed attack system's first component is called the "Interface," and it is the communication/watching module that allows for communication between the agent, the agent's users, and the agent's environment (including other applications) using a variety of interfaces. The second component of the self-directed attack system is the "Memory Module," which contains contextual memory for both short- and long-term usage, as it relates to the activities that the agent is involved with and the data that the agent uses for interaction with prior user systems. The third component of the self-directed attack system is the "Profile Module." The Profile Module contains the definitions for the 'operational framework' of the self-directed attack system, the activities that the self-directed attack system may perform, the operational objectives that the self-directed attack system has been assigned to accomplish, and the behaviors that have been established as predetermined behaviors for the self-directed attack system (Eva, n.d.; Kshetri 2023). The "Planning Module" is comprised of various artificial intelligence techniques that are used by the self-directed attack system to generate multiple-step actionable plans to achieve its goals. The "Action Module" supports agents executing their plans within their onsite environment through the use of external tools and APIs (Evani, n.d.; Boston Consulting Group, 2025).

Several distinct technological methods for developing intelligence and evasion capabilities in these platforms use machine learning and build on top of existing techniques by incorporating malicious intentions. The following is a brief explanation of five of the most relevant technological categories:

- *Deep Neural Networks (DNNs)* are an essential part of developing complex and convoluted decision-making processes that extend beyond the simple Conditional Logic used to create traditional malware (Guembe et al., 2022; Kirat et al., 2018). The greater complexity and difficulty involved in identifying malicious processes from legitimate processes have made it even more challenging for organizations to discern between the two (Guembe et al., 2022; Kirat et al., 2018).
- *Generative Adversarial Networks (GANs)* are a newer form of artificial intelligence that has created new ways for developing intelligent concealment or stealth capabilities. This includes the development of adversarial malware with the intent of escaping detection by machine learning technologies and the creation of undetectable URLs (Anderson et al., 2016; Hu & Tan, 2021; Guembe et al., 2022).
- *Fuzzy Models* allow for the continual construction of malware variants using observed environmental data combined with machine learning (Guembe et al., 2022; Kaloudi & Li, 2020); this is termed "next-generation" malware because it learns continuously from its environment and creates new variants to evade detection (Guembe et al., 2022; Kaloudi & Li, 2020).
- *Reinforcement Learning (RL) and Neural Networks (NNs)* work together to provide behavioral analysis and exploration, providing a means for malware to discover new vulnerabilities and evade web-based authentication systems (Guembe et al., 2022; Petro & Morris, 2017).
- *K-Means Clustering and Gaussian Distribution* Can Help Intelligent Malware to Identify an appropriate time and achieve maximum impact during the attack on targeted computer systems (Chung et al., 2019; Guembe et al., 2022).

Functional Characteristics

In the case of systems that conduct self-directed attacks, the agency of an intelligent AI system determines its operational characteristics since they perform a very high level of autonomous reasoning, adaptation to changes in the surrounding environment, and independence to execute complex tasks (Evani, n.d.; Kshetri, 2025).

Agentic AI systems have agency and can define goals and carry out multistage operations autonomously with minimal input from humans (Kshetri, 2025; Klappholz, 2025). One important functional characteristic is the ongoing cycle of continuous and recursive reasoning (Evani, n.d.; Russell & Norvig, 2021). The agent will continually sense its environment, establish context-based goals for the operation, plan when to attack the target, and execute actions. Then, it integrates the operational outcomes and feedback into the perception phase (Evani, n.d.; Russell & Norvig, 2021). This loop enables real-time learning and optimization (Evani, n.d.).

Due to the autonomy of AI, malware capable of self-learning can intentionally execute pre-defined attack plan s without a human hacker's specialized knowledge (Chung et al., 2019; Guembe et al., 2022). Systems like Deep Locker utilize Deep Neural Networks (DNNs) to disguise their malicious content and only activate the attack once certain target characteristics are present within the surrounding environment, illustrating advanced evasion techniques (Guembeet al., 2022; Kirat et al., 2017).

The system design of such technology allows them to operate successfully in a dynamic environment without having constant oversight from humans (Guembe et al, 2022; Kirat et al., 2017). AI techniques allow for the possibility of executing multi-layered attacks where the self-learning malware would autonomously determine when and how it would deliver its payload within network nodes. Thus, this would negate the requirement for a traditional Command and Control (C2) network (Guembe et al., 2022; Kirat et al., 2017). For example, when it comes to AI-based Distributed Denial of Service (DDoS) attacks, human intervention is no longer necessary because the machines have the ability to adapt their attack vectors without human interference after receiving tactics from the target defenses (Guembeet al., 2022; Kaloudi and Li, 2020).

4.2.3. RQ3: How do AI-Enabled Attacks Exploit Autonomy, Adaptability, and Scalability to Bypass Traditional Cybersecurity Defenses?

AI-driven attacks are more adaptive, scalable, and capable of independent operation compared to most traditional security appliances because they eliminate the need for heavily dependent human involvement and signature-based detection methods that do not allow for the rapid reaction of a defense against such attacks (Guembe et al. 2022, Reddem 2023, Wettstein 2025). By way of their unique adaptive and autonomous characteristics, AI attacks by alternative measurements can completely alter the current state of cyber vulnerabilities and threats (Guembe et al. 2022, Kshetri 2025). Essentially, AI attacks can be looked at as a new trajectory of cyber warfare or a new front/category of warfare (Guembe et al. 2022, Kshetri 2025).

Malicious actors can develop their own set of controls, which is a departure from the traditional model of needing a continually evolving combination of rules and behaviors to define malicious behavior (Guembe et al. 2022, Kshetri 2025). AI-based systems are not limited to this form of simple logic; they require significantly more complex logic to carry out their operations than do conventional targeted attacks based on pre-scripted behavior rules (Guembe et al. 2022, Kirat et al. 2018, Reddem 2023, Wettstein 2025). The degree of complexity of the logic is very high; therefore, it is very difficult for defenders to distinguish between the malicious code used in a malicious attack and the benign processes performed by the system (Guembe et al. 2022).

An agentic AI system can autonomously determine its own goals, analyze/compute, and perform multi-phase operations independently with little input from a human (Kshetri, 2025). This capability enables adversarial actors to conduct highly complex attacks, for instance, by compromising a supercomputing center's cooling systems while making it appear as if there was an unintentional failure (Chung et al., 2019; Guembe et al., 2022). Using AI, malicious programs can auto-schedule the ideal time and approach for deploying their malicious code across a target network (Guembe et al., 2022; Kirat et al., 2018). This has eliminated most of the need for traditional command-and-control methods, thus complicating the detection and halt of such malware through conventional network monitoring (Guembe et al., 2022).

Malware employing AI is also able to monitor current target environments and countermeasures in real-time, allowing it to perpetually change its code to avoid being detected by current traditional antivirus products (Guembe et al., 2022; Reddem, 2023; Wettstein, 2025). The level of complexity seen in malware utilizing AI-powered techniques has allowed such malware to change its characteristics frequently and thus evade both signature- and behavior-based detection by contemporary antivirus products (Babuta et al., 2020; Guembe et al., 2022; Kirat et al., 2018). An example of such evasiveness is epitomized by the deep Locker malware, which hides its payload and only activates when it detects specific targets via their traits (Guembe et al., 2022; Kirat et al., 2018).

Through Machine Learning capabilities (RL) and Fuzzy Logic Systems, next-generation malware will become increasingly sophisticated over time. As both systems adapt to changes, both items can update themselves through a series of fuzzy, variable algorithms and through user experiences; thus, both systems improve themselves continually, based on previous experience. Both malware types share several similarities with traditional malware; this includes auto-propagating/self-replicating/self-erasing/retooling, etc. Cybercriminals are exploiting similar techniques/technologies also used by cybercriminals to evade traditional security measures, but are now employing these techniques and technologies for use against AI-based security solutions, through an adversary approach (Guembe et al., 2022; Kshetri, 2025; Shinde et al., 2024; Walter et al., 2024). There are numerous examples of how an adversarial tactic can impact Entropic AI Detection Systems; examples include Data Poisoning and Pronunciation Error Input, which lead to false identifications and unpredictable system behavior.

The increased level of Automation offered using AI allows Cybercriminals to operate at a speed and volume that far exceeds that of any standard Security Team or all current reactive approaches to Cybercrime (Guembe et al., 2022; Reddem, 2023). The time it takes to complete a cyber-attack is significantly reduced using AI's automation capabilities (Guembe et al., 2022; Reddem, 2023), with an average reduction of 93.1 percent (from 4.6 hours down to 19 minutes) that made human response times to such attacks virtually non-existent (Reddem, 2023). With this tremendous pace of executing cyber-attacks, cybercriminals now can execute significantly greater activity per individual threat actor, leading to a massive 94% increase in attack volume per threat actor (Reddem, 2023).

Utilizing AI-powered capabilities to collect, analyze, and process significant amounts of data (OSINT, social media) allows cybercriminals to utilize more advanced techniques, such as intelligent targeting and vulnerability mapping (Guembe et al., 2022; Kazimierczak et al., 2024; Yamin et al., 2021). AI-Powered tools speed up the beginning phases of the Cyber Kill Chain to the point where they can develop strategies to exploit the Cyber Kill Chain by identifying how cybercriminals can break into their chosen system (Guembe et al., 2022; Kazimierczak et al., 2024).

Generative AI increases the ability of cybercriminals to use deception techniques to enhance their efforts by producing hyper-personalized phishing campaigns and fake media through deepfake technology (Araromi, 2023; Zandi et al., 2024). As a result of utilizing natural language and emotional appeal in their Phishing campaigns, Generative AI increased the success rate of AI phishing attacks over those utilizing traditional methods to 8.7% in 2023, compared to 2.9% for non-AI methods (Reddem, 2023; Zandi et al., 2024).

Even though AI technologies have not been as widely utilized by governmental agencies and organizations, AI has greatly aided the evolution of DDoS attacks, allowing for real-time changes to attack type and vectors, enabling computers to execute highly effective attacks on cybersecurity defenses without human input (Guembe et. Al., 2022).

4.2.4. RQ4: What Limitations Exist in Current Defensive Approaches for Detecting and Mitigating Self-Directed AI-Based Cyberattacks?

Due to weaknesses in technical design within existing conventional security models, adversarial AI pretests of defensive technology, and operational complications relating to AI-based threat ability, the existing methods for detecting and countering self-directing, AI-based cyber-attacks are currently very limited (Guembe et al., 2022; Kshetri, 2025; Reddem, 2023; Vaid, 2023).

The older-generation Cybersecurity Technologies and Techniques are fundamentally incapable of keeping up with the pace of innovation and sophistication of autonomous systems and created by AI (Faheem, 2024; Guembe et al., 2022; Reddem, 2023). Detection systems developed within the traditional model utilize a rule-based and signature-based structure to create a set of 'known attack patterns' (Kumar & Sangwan, 2012; Viegas et al., 2021; Yang et al., 2013). Most detection systems are built within this model, thus making them not effective against the unknown or novel threats or zero-day vulnerabilities (Naiping & Genyuan, 2010; Reddem, 2023; Tiwari et al., 2020).

According to the literature reviewed, on average, traditional Intrusion Detection Systems (IDS) provide approximately 76% effectiveness in the detection of known threats and only approximately 23% effectiveness in the detection of zero-day vulnerabilities (Reddem, 2023). The current state of the cyber defense infrastructure is inadequate and will continue to degrade as it moves toward a contest between machines operating more efficiently as a result of AI and humans relying on their ability to provide labor (Faheem, 2024; Guembe et al., 2022; Reddem, 2023). The rapidity and volume of AI assault operations make detecting and mitigating these assaults with traditional detection and mitigation methods impossible (Guembe et al., 2022; Reddem, 2023).

Subsequently, when organizations begin using AI technologies to defend their systems, they present new vulnerabilities that may be exploited by autonomous offensive AI (Adversarial AI – AAI) (Kshetri, 2025; Shinde et al., 2024; Walter et al., 2024). Consequently, while the usage of Deep Neural Networks (DNN) in defensive AI systems has led to improved performance and capabilities, new vulnerabilities arise during the entire AI Life Cycle (Shinde et al., 2024; Walter et al., 2024). Adversaries can exploit vulnerabilities in the training phase of AI models by introducing malicious examples into the data or changing the characteristics of training data points (Shinde et al., 2024; Walter et al., 2024).

This contamination can cause algorithms to behave incorrectly and result in the compromise of a defensive AI system (Kshetri, 2025). Attackers can craft adversarial inputs and/or create small, unobtrusive changes to data points that would appear to be the same to humans (Adewale, 2022; Shinde et al., 2024). These minor perturbations cause the defensive AI classifier to make incorrect predictions or misclassify threats as benign, demonstrating stealth malware capabilities (Guembe et al., 2022; Shinde et al., 2024).

Model Backdoors further complicate the situation, as adversaries can compromise the model by injecting a backdoor/trigger into the neural network, often during the training phase (Walter et al., 2024). The compromised model appears to function normally until the physical or digital trigger is activated after deployment (Walter et al., 2024). Adversarial examples generated against one machine learning model sometimes have the property of transferability, allowing them to be used effectively to attack a separate, distinct model (Kazimierczak et al., 2024).

Many current defense techniques are highly specialized, capable of addressing only a particular kind of attack (Shinde et al., 2024). Studies show that no single defense strategy has yet been capable of managing all sorts of adversarial attacks (Shinde et al., 2024). Additional general constraints around the ability to see clearly within defense systems, data usage, and the capability to implement advanced artificial intelligence systems create another layer of limitations on the effectiveness of defense mechanisms, besides the targeted forms of attack. Newer generations of AI tools, particularly those using deep learning-type architectures, represent their processes in an

unexplained manner. As described by Guembe et al. (2022), Jameel and Saud (2022), Vaid (2023), and Wettstein (2025), the internal processes or mathematical logic that create the response or outcome of an advanced AI system can be very complex, making it difficult or even..., impossible for a user to verify whether a machine learning model was hacked, or is simply "not working."

The amount of data and quality of that data will significantly impact how effectively and properly an AI-based tool will perform (Shinde et al., 2024; Vaid, 2023), as noted above. Most models rely primarily on past patterns and historical data (Tiwari et al., 2020); they are inherently restricted in their capacity to adapt to novel or emerging threats for which limited training data exists (Tiwari et al., 2020; Vaid, 2023).

Security AI defense systems generate a lot of false positives, which result in the security team mistaking normal activity for potentially problematic activity (Naiping and Genyuan 2010, Tiwari et al. 2020, and Vaid 2023). This over-burdening with incorrectly identifying threats creates "alert fatigue" on the part of the security personnel and causes them to become complacent about legitimate threats (Tiwari et al. 2020, and Vaid 2023). It has also been shown by studies that there is still not enough knowledge about how AI technologies operate in the real world, since the majority of studies are based solely upon using specific techniques tested in controlled laboratory conditions (Ofusori et al. 2024, Sommer and Paxson 2010, and Walter et al. 2024).

The rapid growth of AI has outpaced the workforce; many organizations struggle to find cybersecurity specialists qualified to develop, implement, and maintain complex AI systems (Tambe et al., 2020; Zandi et al., 2024). Additionally, the high costs associated with acquiring and maintaining AI technologies pose a financial barrier to widespread adoption (Obyed et al., 2024).

4.3. Discussion

This study positions Artificial Intelligence (AI) as a transformative force in cybersecurity, marking a decisive shift from static, reactive defenses toward adaptive, learning-driven security ecosystems. Instead of reiterating established claims that AI enhances detection accuracy, the discussion emphasizes how and why AI fundamentally reshapes cyber defense logic, operational tempo, and strategic posture. The findings synthesize prior research into a coherent framework that explains AI-enabled cybersecurity as a move from rule enforcement to behavioral inference and anticipatory risk management.

A key insight from this review is that AI-driven cybersecurity benefits not just from computational speed but from its ability to model complex, evolving system behavior. Traditional intrusion detection mechanisms are inherently limited by their reliance on predefined signatures and thresholds, which restrict their effectiveness against zero-day exploits, polymorphic malware, and insider threats. In contrast, AI-based systems establish behavioral baselines across users, endpoints, and networks, allowing them to detect statistically significant deviations even without known attack indicators.

This behavioral focus marks a qualitative shift from previous methods and explains the consistently superior performance of AI-enhanced detection systems observed across domains. The discussion further highlights the role of deep learning architecture in extending detection beyond surface-level anomalies to multi-stage and low-and-slow attack patterns. By learning temporal dependencies and structural relationships within security telemetry, these models reveal attack strategies that unfold incrementally and evade conventional monitoring. Importantly, this capability addresses a key gap in earlier cybersecurity literature, which often treated attacks as discrete events rather than dynamic processes unfolding over time.

Beyond detection, the review highlights predictive cyber threat intelligence as a critical but under-theorized contribution of AI to cybersecurity. While traditional threat intelligence frameworks are largely retrospective, AI-

driven systems operate prospectively, inferring attacker intent and emerging risks from both structured telemetry and unstructured information sources. This shift redefines threat intelligence from just an informational artifact into an operational input for real-time decision-making. In high-risk sectors, such predictive capability enables continuous risk scoring and adaptive control enforcement, aligning security responses with contextual threat levels rather than static policy rules.

The emergence of agentic AI marks a further escalation in cybersecurity maturity. Unlike earlier automation tools, agentic systems demonstrate goal-directed behavior, recursive reasoning, and coordinated action across security functions. The discussion emphasizes that the main contribution of agentic AI is not full autonomy itself, but the reduction of decision cycles in environments where human response alone cannot scale. By automating alert triage, contextual enrichment, and routine containment actions, agentic AI significantly reduces operational friction and helps mitigate analyst fatigue, two persistent challenges highlighted in the cybersecurity literature.

Crucially, this review emphasizes that effective AI-driven cybersecurity does not eliminate human involvement but redefines it. Human–machine teaming models emerge as a dominant paradigm, wherein AI systems manage complexity and speed while human operators retain strategic oversight and ethical accountability. This synthesis advances prior work by resolving the often-framed dichotomy between automation and control, demonstrating instead that AI augments human decision-making when appropriately governed.

At the architectural level, the discussion links AI adoption to broader strategic shifts such as zero-trust security models and continuous authentication. AI enables these paradigms by dynamically recalculating trust based on behavioral and contextual signals, thereby operationalizing concepts that were previously difficult to implement at scale. The integration of explainable AI further addresses a critical adoption barrier by improving transparency, auditability, and regulatory compliance. This is especially important in regulated sectors, where opaque decision-making has historically constrained the deployment of advanced analytics.

The review also acknowledges the continued importance of deterministic security mechanisms, arguing that probabilistic AI and rule-based controls are complementary rather than competing approaches. Deterministic methods provide immediate response triggers for unequivocal violations, while AI excels in probabilistic inference under uncertainty. This hybrid perspective advances the literature by proposing a balanced defense model that reduces false positives without sacrificing decisiveness.

Sector-specific analysis reinforces the generalizability of these findings. Across finance, healthcare, and defense, AI-driven cybersecurity consistently demonstrates enhanced detection, faster response, and improved resilience. However, the discussion deliberately avoids technological determinism, acknowledging that AI introduces new risks, including model drift, adversarial manipulation, and governance complexity. These challenges underscore the necessity of robust lifecycle management, continuous validation, and ethical oversight—areas that warrant further empirical investigation.

As stated previously, malicious users will continue to exploit AI-based technologies to assist them in making better choices. They are now able to gain contextual information from large amounts of data through the method of emulation, using trusted properties associated with the cyber environment or by attacking known flaws; as a result, these new AI-powered attack methods will be inherently more challenging for victims to detect and defend against than current attack methods, and will likely have a significantly greater impact on the overall state of cybersecurity.

In addition to being modified and evolving as they interact with the environment, these continuously evolving forms of cyberattacks will also create new challenges for organizations trying to defend against them. As a result of these new forms of attacks, continuously learning and adapting to their surroundings, organizations will have to adapt to the rapid growth in technology and the increasing difficulty of finding these attacks.

While organizations currently rely heavily on the ability of employees to identify new threats through traditional means, AI will continue to increase their capabilities. This means that organizations will soon become more reliant on technology, rather than on the ability of humans to identify and respond to threats. Consequently, organizations will be ill-prepared for the new forms of AI-based cyber threats they will inevitably encounter, which will lead to more frequent attacks and will require organizations to develop more sophisticated methods of protecting themselves from such attacks.

Utilizing AI to defend against other forms of AI-based threats will, therefore, enable researchers, organisations, and Governments to develop more advanced methods for combating AI threats in the future. Consequently, the highest level of current preparatory activity consists of developing a Cyber Defence System that gives the user as much confidence as possible against any potential false positive or negative threat assessment.

In summary, this discussion advances the cybersecurity literature by synthesizing fragmented findings into an integrated conceptual understanding of AI-enabled defense. Instead of viewing AI as just a collection of tools, it presents AI as a structural enabler of adaptive, predictive, and resilient cybersecurity architectures. This perspective responds to reviewer concerns about depth and contribution by clarifying how AI changes the logic, pace, and governance of cybersecurity, while also highlighting critical limitations that influence future research and practice.

4.4. Legal, Ethical, and Policy Analysis

The integration of legal, ethical, and policy frameworks is essential for managing the "dual-faced entity" of Artificial Intelligence (AI), which acts as both a powerful defense mechanism and a sophisticated tool for cybercrime (Adewale, 2022; Vaid, 2023). As AI outpaces traditional governance, a multi-layered approach is necessary to address the shifting boundaries of accountability, privacy, and national security (Araromi, 2023; Jameel & Saud, 2022).

4.4.1. Legal Analysis: Jurisdictional Gaps and Liability

Current legal regimes are struggling to keep up with the autonomous and unpredictable nature of agentic AI systems.

- *Insufficiency of Cybercrime Law:* Traditional frameworks, such as the Budapest Convention, are often "human-centric" and do not adequately address novel AI crimes like deepfakes, data poisoning, and the autonomous commission of crimes (Araromi, 2023; Jameel & Saud, 2022). Legislative gaps occur because most laws require "intentional acts" by a human, whereas AI can launch attacks based on learned behaviors that may not have been specifically programmed by the user (Araromi, 2023; Kazimierczak et al., 2024).
- *The Problem of Attribution and Liability:* Assigning criminal responsibility for autonomous AI actions remains a challenge (Araromi, 2023; Wettstein, 2025). Legal scholars suggest models such as "perpetrator-via-another," where the programmer is liable for using the AI as a tool, or strict liability, where the "keeper" of a dangerous autonomous system is responsible for its actions regardless of direct intent (Araromi, 2023).
- *Regulatory Milestones:* The EU AI Act is the first comprehensive effort to harmonize AI rules, using a risk-based approach to categorize systems from "minimal" to "unacceptable" risk (Jameel & Saud, 2022; West, 2020). Similarly, regional laws like Thailand's PDPA are being adapted to regulate AI-driven data processing and enforce user consent (LEESA-NGUANSUK, 2019; Wongrass, 2023).

4.4.2. Ethical Analysis: Bias, Transparency, and Human Agency

The ethical aspect of AI is increasingly centered on the "black box" problem, where the reasoning behind automated decisions is cognitively inaccessible to humans (Evani, n.d.; Rugge, 2020; Shinde et al., 2024).

- Explainable AI (XAI) and Accountability: In high-stakes fields like healthcare and defense, the lack of transparency complicates forensic attribution and erodes trust (Shinde et al., 2024; Wettstein, 2025). The principle of "meaningful human control" is crucial to ensure that life-and-death decisions—such as those made by autonomous weapons or medical diagnostic tools—are never fully delegated to software (Benanti, 2020; Guttierri, 2025; Walter et al., 2024).
- Algorithmic Bias: AI systems often replicate or amplify societal prejudices found in their training data (Uddin et al., 2025; West, 2020). For instance, voice assistants and bots have been criticized for reinforcing gender stereotypes by using female-sounding voices for submissive roles or by responding evasively to harassment (Chin & Robison, 2020).
- Human-Centric Paradox: While Industry 5.0 emphasizes human-AI collaboration, the increasing autonomy of GenAI can supersede human agency (Wasi et al., 2025). Ethical governance must balance the pursuit of efficiency with the preservation of human dignity and social equity (Benanti, 2020; Youvan, 2024).

4.4.3. Policy Analysis: Strategic Competition and Global Governance

AI is no longer just a technology; it is a strategic national asset central to global power hierarchies and sovereignty (Rugge, 2020; Wasi et al., 2025).

- Race for Supremacy: The U.S., China, and Russia are engaged in an arms race for AI leadership, which is driving the geopolitical fragmentation of the internet and ICT supply chains (Jameel & Saud, 2022; Rugge, 2020; Wasi et al., 2025). This leads to "technological surprise" risks, where nations may deploy unsafe or untested systems to maintain a first-mover advantage (Rugge, 2020; Scharre, 2019).
- Evolving Defensive Paradigms: Military policy is shifting from "deterrence" (punishment) to "persistent engagement" and "defending forward" (Sullivan, 2025). This requires proactive, real-time maneuvers in the cyber domain to disrupt adversary infrastructure before attacks reach full maturity (Fogarty, 2025; Guttierri, 2025).
- Data Sovereignty vs Localization: Policies favoring data localization keep data within borders, enhancing national control but hindering the global collaboration necessary for AI innovation (Wasi et al., 2025). Emerging policies must find a middle ground, such as using federated learning to analyze data without moving it across jurisdictions (Kouklaras et al., 2025; Shinde et al., 2024).
- Workforce and Literacy: Policies must address the cybersecurity skills gap, eestimated at nearly 4 million professionals, and promote ethical literacy to prepare citizens and employees for a world saturated with AI content and decision-making (Araromi, 2023; Kshetri, 2025; Youvan, 2024).

5. Conclusion

Cybercriminals are constantly improving their attack methods by utilizing various forms of AI technology. This research provides insight into how AI can enhance the capabilities of attackers to carry out large-scale attacks more rapidly and on a wider scale. It also reviews prior research related to cyberattacks using AI, independent systems capable of conducting attacks, and the negative effects of AI-based cyberattacks on society and the economy. The analysis found that 56% of AI-based attack methods occur during access and penetration phases, 12% happen during exploitation phases, 11% occur in reconnaissancee, and 9% take place in the delivery phase. CNNs are the most common methods used in this area (5) to illustrate access and penetration attacks.

According to this study, 63% of all studies on AI-related cybersecurity have focused on implementation and evaluation, 25% of research has presented frameworks, and 12% of research has focused on how to attack using AI.

Most traditional cybersecurity solutions are ineffective at detecting or preventing attacks from AI because of the rapid pace, complex decision-making, and various types of AI-based threats.

For organizations and security teams, it is crucial to quickly adapt their strategies for using AI to defend against these advanced attacks. This report also emphasizes that it is essential for the Security Research community, Nation-States, and Cybersecurity Experts to collaborate in developing and investing in new innovative countermeasures against AI cybersecurity threats, and to demonstrate that AI can also be used offensively to carry out cyber-attacks. The principles used to create advanced detection logic will also be developed with trustworthy AI in the future.

References

[1] Rakan A. Alsowail, and Taher Al-Shehari, "Techniques and Countermeasures for Preventing Insider Threats," *PeerJ Computer Science*, vol. 8, pp. 1-37, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] W. Elghazel et al., "Dependability of Wireless Sensor Networks for Industrial Prognostics and Health Management," *Computers in Industry*, vol. 68, pp. 1-15, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[3] Muhammad Imran et al., "A Performance Overview of Machine Learning-based Defense Strategies for Advanced Persistent Threats in Industrial Control Systems," *Computers and Security*, vol. 134, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Manpreet Singh, The Rise of AI-Driven Cybersecurity: A New Era of Defense and Offense - ET Edge Insights, 2024. [Online]. Available: https://etedge-insights.com/technology/artificial-intelligence/the-rise-of-ai-driven-cybersecurity-a-new-era-of-defense-and-offense/.

[5] Abel Yeboah-Ofori, and Francisca Afua Opoku-Boateng, "Mitigating Cybercrimes in an Evolving Organizational Landscape," *Continuity and Resilience Review*, vol. 5, no. 1, pp. 53-78, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] Temitope Adewale, "Artificial Intelligence in Cybercrime: Unveiling the Emerging Landscape of Intelligent Threats," 2022. [Google Scholar]

[7] Marcus Ayodeji Araromi, Artificial Intelligence-Enabled Cyber-Criminality: Is the Current Cybercrime Legal Regime Sufficient?, University of Ibadan, Ibadan, Nigeria, 2023. [Online]. Available: https://www.cavendish.ac.ug/wp-content/uploads/2025/04/4-Artificial-Intelligence-Enabled-Cyber-Criminality-Is-the-Current-Cybercrime-Legal-Regime-sufficient.pdf

[8] Kousik Barik et al., "Cybersecurity Deep: Approaches, Attacks, Dataset, and Comparative Study," *Applied Artificial Intelligence*, vol. 36, no. 1, pp. 1-24, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Thomas A. Campbell, *Cybersecurity in AI National Strategies*, AI in the Age of Cyber-Disorder Actors, Trends, and Prospects, Ledizioni Ledi Publishing, pp. 56-62, 2020. [Online]. Available: https://www.ispionline.it/sites/default/files/pubblicazioni/ispi_report_ai_in_the_age_of_cyber-disorder_2020.pdf

[10] Joe Burton et al., AI and Serious Online Crime, Centre for Emerging Technology and Security, The Alan Turing Institute, 2025. [Online]. Available: https://cetas.turing.ac.uk/publications/ai-and-serious-online-crime

[11] Thomas A. Campbell, *Cybersecurity in AI National Strategies*, AI in the Age of Cyber-Disorder Actors, Trends, and Prospects, Ledizioni Ledi Publishing, pp. 56-62, 2020. [Online]. Available: https://www.ispionline.it/sites/default/files/pubblicazioni/ispi_report_ai_in_the_age_of_cyber-disorder_2020.pdf

[12] Caitlin Chin-Rothmann, and Mishaela Robison, *How AI Bots and Voice Assistants Reinforce Gender Bias*, AI in the Age of Cyber-Disorder Actors, Trends, and Prospects, Ledizioni Ledi Publishing, pp. 82-104, 2020. [Online]. Available: https://www.ispionline.it/sites/default/files/pubblicazioni/ispi_report_ai_in_the_age_of_cyber-disorder_2020.pdf

[13] Samuele Dominioni, *Panopticon 2.0? AI Enabled Surveillance Practices in Authoritarian Reg*, AI in the Age of Cyber-Disorder Actors, Trends, and Prospects, Ledizioni Ledi Publishing, pp. 63-81, 2020. [Online]. Available: https://www.ispionline.it/sites/default/files/pubblicazioni/ispi_report_ai_in_the_age_of_cyber-disorder_2020.pdf

[14] Foluke Ekundayo et al., "Predictive Analytics for Cyber Threat Intelligence in Fintech using Big Data and Machine Learning," *International Journal of Research Publication and Reviews*, vol. 5, no. 11, pp. 5934-5948, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Pavan Kumar Evani, "Agentic AI Security: A Conrol Framework for Autonomous Decision-Making Systems," *SSRN*, pp. 1-28, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[16] Scott C. Fogarty, "The Sword of Damocles: A Cybersecurity Paradigm Shift for the Defense of Critical Infrastructure," *The Cyber Defense Review*, vol. 10, no. 1, pp. 29-39, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[17] Taban Habibu, and Ayo P. Julius, "Cybersecurity in the Internet of Things (IoT) – Review," *DS Journal of Cyber Security*, vol. 3, no. 3, pp. 15-38, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[18] Leonel Garciga, and Deborah S. Karagosian, "Conversation with the U.S. Army Chief Information Officer," *The Cyber Defense Review*, vol. 10, no. 1, pp. 17-28, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[19] A. Shaji George, "Riding the AI Waves: An Analysis of Artificial Intelligence's Evolving Role in Combating Cyber Threats," *Partners Universal International Innovation Journal (PUIIJ)*, vol. 2, no. 1, pp. 39-50, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[20] Yasmine Ghazlane, Maha Gmira, and Hicham Medromi, "Development of a Vision-based Anti-Drone Identification Friend or Foe Model to Recognize Birds and Drones using Deep Learning," *Applied Artificial Intelligence*, vol. 38, no. 1, pp. 1-30, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Blessing Guembe et al., "The Emerging Threat of AI-Driven Cyber Attacks: A Review," *Applied Artificial Intelligence*, vol. 36, no. 1, pp. 1-34, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[22] Karen Guttieri, "Fighting through Disruption: Reframing Cyber Resilience for Power Projection and Strategic Credibility," *The Cyber Defense Review*, vol. 10, no. 1, pp. 93-114, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[23] Tanzeela Jameel, and Adam Saud, "Policies of Artificial Intelligence in the EU: Learning Curve from the UK and China?," *Journal of European Studies*, vol. 38, no. 2, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[24] Fnu Jimmy, "Emerging Threats: The Latest Cybersecurity Risks and the Role of Artificial Intelligence in Enhancing Cybersecurity Defenses," *International Journal of Scientific Research and Management (IJSRM)*, vol. 9, no. 2, pp. 564-574, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Mateusz Kazimierczak et al., "Impact of AI on the Cyber Kill Chain: A Systematic Review," *Heliyon*, vol. 10, no. 4, pp. 1-21, 2024. [Google Scholar] [Publisher Link]

[26] Habib Ullah Khan, Muhammad Zain Malik, and Shah Nazir, "Identifying the AI-Based Solutions Proposed for Restricting Money Laundering in Financial Sectors: Systematic Mapping," *Applied Artificial Intelligence*, vol. 38, no. 1, pp. 1-31, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] Obyed Ullah Khan et al., "The Future of Cybersecurity: Leveraging Artificial Intelligence to Combat Evolving Threats and Enhance Digital Defense Strategies," *Journal of Computational Analysis and Applications*, vol. 33, no. 8, pp. 776-787, 2024. [Google Scholar] [Publisher Link]

[28] Christos Koukaras et al., "AI-Driven Telecommunications for Smart Classrooms: Transforming Education through Personalized Learning and Secure Networks," *Telecom*, vol. 6, no. 2, pp. 1-26, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[29] Nir Kshetri, "Transforming Cybersecurity with Agentic AI to Combat Emerging Cyber Threats," *Telecommunications Policy*, vol. 49, no. 6, pp. 1-12, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[30] Lizzy Ofusori, Tebogo Bokaba, and Siyabonga Mhlongo, "Artificial Intelligence in Cybersecurity: A Comprehensive Review and Future Direction," *Applied Artificial Intelligence*, vol. 38, no. 1, pp. 1-46, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[31] C. Rajathi, and Rukmani Panjanathan, "A Two-Phase Feature Selection Framework for Intrusion Detection System: Balancing Relevance and Computational Efficiency (2P-FSID)," *Applied Artificial Intelligence*, vol. 39, no. 1, pp. 1-29, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[32] Jakiur Rahman et al., "A Generalized and Robust Nonlinear Approach based on Machine Learning for Intrusion Detection," *Applied Artificial Intelligence*, vol. 38, no. 1, pp. 1-34, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[33] Poli Reddy Reddem, "The Rise of AI-Powered Cybercrime: A Data-Driven Analysis of Emerging Threats," *International Journal for Multidisciplinary Research (IJFMR)*, vol. 6, no. 6, 2024. [CrossRef] [Publisher Link]

[34] Jeth B. Rey, "The Battlefield is Not 'Over There'- It is Here, 24/7," *The Cyber Defense Review*, vol. 10, no. 1, pp. 5-15, 2025. [Google Scholar] [Publisher Link]

[35] Fabio Rugge, *AI in a Contested Cyberspace*, AI in the Age of Cyber-Disorder Actors, Trends, and Prospects, Ledizioni Ledi Publishing, pp. 12-55, 2020. [Online]. Available: https://www.ispionline.it/sites/default/files/pubblicazioni/ispi_report_ai_in_the_age_of_cyber-disorder_2020.pdf

[36] Abdu Salam et al., "Deep Learning Techniques for Web-Based Attack Detection in Industry 5.0: A Novel Approach," *Technologies*, vol. 11, no. 4, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[37] Krutika Sawant et al., "A Study of AI in Banking System," *Korea Review of International Studies*, vol. 16, special no. 6, 2023. [Google Scholar] [Publisher Link]

[38] Rucha Shinde et al., "Securing AI-Based Healthcare Systems using Blockchain Technology: A State-of-the-Art Systematic Literature Review and Future Research Directions," *Transactions on Emerging Telecommunications Technologies*, vol. 35, no. 1, pp. 1-48, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[39] Wadim Strielkowski et al., "AI-Driven Adaptive Learning for Sustainable Educational Transformation," *Sustainable Development*, vol. 33, no. 2, pp. 1921-1947, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[40] Scott Sullivan, "Toward Clarity in Cyber's "Fog of law"," *The Cyber Defense Review*, vol. 10, no. 1, pp. 59-71, 2025. [Google Scholar] [Publisher Link]

[41] Cadet Brandon Tran, "Southeast Asia: Where Facebook is the Internet," *The Cyber Defense Review*, vol. 10, no. 1, pp. 41-58, 2025. [Google Scholar] [Publisher Link]

[42] Mueen Uddin et al., "A Critical Analysis of Generative AI: Challenges, Opportunities, and Future Research Directions," *Archives of Computational Methods in Engineering*, pp. 1-31, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[43] Abbos Utkirov, "Artificial Intelligence Impact on Higher Education Quality and Efficiency," vol. 4, no. 9, pp. 1-27, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[44] Sanjay Vaid, "Impact Assessment of Artificial Intelligence on Cybersecurity: A Review of the Existing Literature," *Journal of WTO and International Business*, vol. 25, no. 4, pp. 13-22, 2023. [Publisher Link]

[45] John Villasenor, *How To Deal with AI Enabled Disinformation?*, AI in the Age of Cyber-Disorder Actors, Trends, and Prospects, Ledizioni Ledi Publishing, pp. 105-117, 2020. [Online]. Available: https://www.ispionline.it/sites/default/files/pubblicazioni/ispi_report_ai_in_the_age_of_cyber-disorder_2020.pdf

[46] Mathew J. Walter, Aaron Barrett, and Kimberly Tam, "A Red Teaming Framework for Securing AI in Maritime Autonomous Systems," *Applied Artificial Intelligence*, vol. 38, no. 1, pp. 1-36, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[47] Azmine Toushik Wasi et al., "Generative AI as a Geopolitical Factor in Industry 5.0: Sovereignty, Access, and Control," *arXiv Preprint*, pp. 1-30, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[48] Darrell M. West, *AI Revolution: Building Responsible Behav*, AI in the Age of Cyber-Disorder Actors, Trends, and Prospects, Ledizioni Ledi Publishing, pp. 118-133, 2020. [Online]. Available: https://www.ispionline.it/sites/default/files/pubblicazioni/ispi_report_ai_in_the_age_of_cyber-disorder_2020.pdf

[49] Benjamin Wettstein, "*Policy and Technical Recommendations for Integrating Autonomy into Military Offensive Cyberspace Operations*," Master's Thesis, Massachusetts Institute of Technology, 2025. [Google Scholar] [Publisher Link]

[50] Piriyapong Wongras, "*A Study of Factors Influencing Employees' Acceptance of Artificial Intelligence Technology in Recruitment*," Master's Independent Study, Thammasat University, 2023. [Google Scholar] [Publisher Link]

[51] Yuqing Xing et al., "Waveforms Eavesdropping Prevention Framework: The Case of Classification of EPG Waveforms of Aphid Utilizing Wavelet Kernel Extreme Learning Machine," *Applied Artificial Intelligence*, vol. 37, no. 1, pp. 1-29, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[52] Xinzhu Yan, "Research on Financial Field Integrating Artificial Intelligence: Application Basis, Case Analysis, and SVR Model-Based Overnight," *Applied Artificial Intelligence*, vol. 37, no. 1, pp. 1-26, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[53] Wenlong Yi et al., "Design and Implementation of a Blockchain-Based Self-Directed Learning Process Evaluation Traceability Platform," *Computer Tools in Education*, no. 4, pp. 70-81, 2024. [Google Scholar] [Publisher Link]

[54] Muhammad Younas, Dina Abdel Salam El-Dakhs, and Yicun Jiang, "A Comprehensive Systematic Review of AI-Driven Approaches to Self-Directed Learning," *IEEE Access*, vol. 13, pp. 38387-38403, 2025. [CrossRef] [Google Scholar] [Publisher Link]