International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

From Phishing Bots to Autonomous Agents: How AI Is Transforming Cybersecurity Threats and Defenses

Priyansh Vats

Guru Gobind Singh Indraprastha University

Abstract

AI technology is revolutionizing the field of cybersecurity by altering not only the attack methodologies but also the defense tools. The earlier kinds of cyber attacks were dependent upon general phishing emails, malware scripts, credential hacking, ransomware, and exploiting the already existing vulnerabilities. However, the emergence of new types of AI such as generative AI and agentic AI has changed the dynamics of the attacks. Now AI can be used by attackers for generating persuasive phishing attacks, automated social engineering, impersonating identities through synthetic media, faster reconnaissance, malware creation, and exploiting the human element. Meanwhile, AI technology is becoming an integral part of cyber defense in terms of anomaly detection, log analysis, threat intelligence, prioritization of vulnerabilities, and incident responses. In this paper, a review of secondary sources of information will be conducted with regard to governmental agencies, cybersecurity companies, technical frameworks, and academic papers. The purpose of this study is to explain how AI is revolutionizing the cybersecurity process and to claim that AI is transforming cybersecurity in three stages, starting from deception through acceleration to an agentic attack surface. The paper concludes that cybersecurity must shift from static detection and isolated technical controls toward identity-centered defense, human oversight, secure AI development, agentic threat modeling, and collaborative incident reporting.

Keywords: artificial intelligence, cybersecurity, phishing, generative AI, agentic AI, ransomware, social engineering, AI governance, cyber defense

Chapter 1: Introduction

Cybersecurity has always evolved alongside technological change. Email produced phishing; cloud computing produced new identity and configuration risks; mobile technology expanded the attack surface; and digital supply chains created dependencies that attackers could exploit indirectly. Artificial intelligence now represents the next major shift. Unlike previous technologies, AI does not merely add another system to be secured. It changes the speed, scale, and sophistication of cyber activity itself.

However, the present discussion about AI and cybersecurity is oversimplified in many ways. On one hand, AI is considered to be mostly a means of defense that detects threats faster than humans can. On the other hand, AI is seen as a dangerous innovation that allows cybercriminals to perform their tasks. The problem with these two perspectives is that they miss the point about the duality of the innovation, which means that the same technology can be used by both defenders and attackers. According to Microsoft's Digital Defense Report 2025, AI becomes necessary for defending against AI-powered attacks and also is a

potential target because adversaries are currently using generative AI for social engineering, discovering vulnerabilities, lateral movement, and evading security controls (Microsoft Threat Intelligence, 2025, pp. 33, 36, 52–53). At the same time, according to the M-Trends 2026 report by Mandiant, attackers start using AI to improve their productivity during reconnaissance, social engineering, and malware creation processes, but still, they have mostly human-caused intrusions.

It's a distinction worth noting. The issue isn't that AI suddenly supersedes all traditional cybersecurity measures. Instead, the issue is that AI allows attacks currently in use to become more effective while at the same time opening up new types of vulnerabilities. A phishing email is nothing new. But AI-generated phishing that is personalized, multilingual, and capable of emotional manipulation is a step above. Malware isn't anything new either. But when malware development gets assistance from AI as well as the ability to evade and adapt to defense mechanisms, that's a step above. Autonomous systems aren't entirely new either. But when autonomous systems get upgraded to agentic AI and are able to make use of tools, memory, API interactions, and human oversight, that's a step above.

This paper therefore studies the transformation of cybersecurity from "phishing bots" to "autonomous agents." It asks, "How is AI changing cyber threats and defenses? What risks emerge when AI is used not only to generate content but also to act across digital systems? And how should organizations respond?

Research Aim

This paper aims to examine how artificial intelligence is transforming cybersecurity threats and defenses, from AI-enabled phishing and social engineering to the emergence of autonomous agentic AI systems as new attack surfaces.

Research Questions

1. How is AI enhancing traditional attacks on cybersecurity, including phishing, social engineering, ransomware, and malware?
2. How does an increase in agentic and autonomous AI create new risks in cybersecurity?
3. How are cybersecurity defenses evolving due to the threat posed by AI?
4. Which framework is currently available for addressing AI risk in cybersecurity?

Research Objectives

1. To analyze the role of AI in strengthening cyberattacks.
2. To study the cybersecurity risks of agentic AI systems.
3. To compare traditional cybersecurity defenses with AI-powered defense mechanisms.
4. To evaluate current mitigation frameworks such as OWASP, CISA, NIST, MITRE, NCSC, Microsoft, ENISA, and Mandiant.

Research Gap

The existing body of research on artificial intelligence in relation to cybersecurity mostly divides into three distinct categories. The first category of research is concerned with artificial intelligence as an attacker's instrument, in particular, as applied to phishing, social engineering, malware creation, ransomware assistance, and fraud through deepfakes. The Microsoft reports, ENISA reports, Mandiant reports, and the results of the surveys conducted by academics are good examples of the existing body of evidence demonstrating how generative AI may be used to amplify cyberattacks in terms of their scale, personalization, and credibility (Microsoft Threat Intelligence, 2025, pp. 33, 36, 52–53; ENISA, 2025, pp. 15–16; Mandiant, 2026, pp. 11–13; Falade, 2023, p. 5). The third category focuses on securing AI systems themselves, especially through frameworks such as the NIST AI Risk Management Framework, NIST Generative AI Profile, OWASP Top 10 for LLM Applications, OWASP Agentic AI Threats and

Mitigations, and the NCSC Guidelines for Secure AI System Development (NIST, 2023, pp. 20–32; NIST, 2024, pp. 10–12; OWASP, 2024, pp. 22–28; OWASP, 2025, pp. 23–41; NCSC et al., 2023, p. 8).

Nevertheless, a significant amount of the literature that exists in this regard considers these areas independently. The conversation about phishing attacks or AI-based attacks does not go far enough towards agentic AI. Likewise, the technical frameworks about securing AI systems tend to limit themselves to model, application, or lifecycle risks but do not make the connection between these and the larger issue of how AI is transforming cybercrime through its ongoing evolution from improved deception and accelerated cyber operations to agentic digital entities.

This research paper fills this gap through the development of a three-step process: AI as a deception amplifier; AI as an operational amplifier; and agentic AI as an autonomous attack surface. It links classic cyber threats like phishing and ransomware with agentic threats such as memory poisoning, excessive agency, misuse of tools, abuse of privileges, and overload of human-in-the-loop. In this sense, the research paper makes a significant contribution to the conceptualization of how AI is changing cyber threats and defenses.

Chapter 2: Research Methodology

The current study relies on a qualitative secondary research method. The research is conducted through a systematic review of cybersecurity threat reports, AI governance frameworks, technical security information, and academic literature published during the years 2023–2026. Sources used in the study were selected because they provide information regarding any of the following topics: AI-powered social engineering, AI-powered cybercrime, threats to security in generative AI, agentic AI security threats, incident reporting regarding AI, and AI risk governance.

The sources used can be divided into three groups. First, threat intelligence reports from Microsoft, ENISA, and Mandiant were utilized in order to learn about actual attack patterns such as phishing, vishing, ransomware, ClickFix, deepfakes, and the use of AI by attackers. Second, technical and governance frameworks by OWASP, CISA, NIST, and NCSC were utilized in order to determine potential risks and countermeasures regarding LLMs, generative AI, and agentic AI. Third, academic papers were used to analyze broader concepts such as AI-powered cybercrime, the balance between attack and defense, malware, and critical infrastructure threats.

The methodology is interpretive rather than statistical. The aim is not to calculate the exact global frequency of AI-enabled cyberattacks, since incident reporting remains inconsistent and many AI-related attacks may not be publicly disclosed. Instead, the paper synthesizes patterns across credible sources to identify how AI is changing cyber threats and defenses. Evidence is organized using the proposed three-stage framework: AI as a deception multiplier, AI as an operational accelerator, and agentic AI as an autonomous attack surface.

Chapter 3: Literature Review

Mandiant. (2026). M-Trends 2026 Special Report. Incident-response and threat intelligence report. Real-world intrusion trends, ransomware, infection vectors, and vishing. Exploits were the most common initial infection vector in 2025 at 32% of identified cases; voice phishing represented 13%; email phishing declined as an initial vector but remains widely used (pp. 11–13). Ransomware-related intrusions accounted for 13% of Mandiant investigations in 2025 (p. 29). Useful for grounding the paper in actual incident-response data rather than speculation.

CISA. (2025). JCDC AI Cybersecurity Collaboration Playbook. Government cybersecurity playbook. AI incident reporting, vulnerability coordination, and information sharing. AI incident reporting should include affected models, datasets, APIs, libraries, agentic/copilot platforms, third-party systems, and AI application access details (p. 14). Includes a prompt-injection incident example involving MathGPT (p. 21). Supports the argument that AI incidents require broader reporting than traditional cyber incidents.

Lohn, A. J. (2025). The Impact of AI on the Cyber Offense-Defense Balance and the Character of Cyber Conflict. Policy / academic analysis. AI and cyber offense-defense balance. Argues that AI may benefit both attackers and defenders, but its impact depends on capability level, access to expertise, and how organizations integrate AI into operations (p. 29). Useful for the discussion section, especially to avoid a simplistic "AI only helps attackers" argument.

ENISA. (2025). ENISA Threat Landscape 2025. Government / institutional threat report. European cyber threat landscape, ransomware, social engineering, and malicious AI use. Threat groups are using commercial LLMs and diverted/jailbroken models such as WormGPT, EscapeGPT, and FraudGPT to automate social engineering and accelerate malicious tool development (p. 16). Supports the argument that AI is being integrated into real-world cybercrime ecosystems.

Microsoft Threat Intelligence. (2025). Microsoft Digital Defense Report 2025. Industry threat intelligence report. AI-era cyber threats, fraud, identity attacks, ransomware, deepfakes, and AI defense. AI is being used to automate phishing, generate synthetic identities, create deepfakes, and support social engineering. Microsoft also identifies ClickFix as a major initial-access trend that bypasses traditional phishing defenses (pp. 33, 36, 52–53). Core source for the "phishing bots" and AI-enabled deception sections.

OWASP. (2025). Agentic AI: Threats and Mitigations. Technical security framework. Agentic AI risks and mitigations. Agentic systems introduce risks such as memory poisoning, tool misuse, excessive agency, identity compromise, multi-agent manipulation, and human-in-the-loop overload (pp. 23–41). Most important source for the "autonomous agents" section.

NIST. (2024). AI Risk Management Framework: Generative AI Profile. Government AI governance profile. Generative AI-specific risks. Notes that generative AI can lower barriers for offensive cyber capabilities and expand the attack surface through prompt injection, data poisoning, malware, hacking, and phishing risks (p. 10). Bridges general AI governance with generative AI cybersecurity risks.

OWASP. (2024). OWASP Top 10 for LLM Applications 2025. Technical security framework. LLM application vulnerabilities. Identifies prompt injection, sensitive information disclosure, supply-chain vulnerabilities, data/model poisoning, improper output handling, excessive agency, system prompt leakage, vector weaknesses, misinformation, and unbounded consumption (pp. 3–5, 22–28). Provides technical vocabulary for LLM-specific cybersecurity risks.

Falade, P. V. (2023). Decoding the Threat Landscape: ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks. Academic paper. AI-enabled social engineering. FraudGPT and Worm GPT are discussed as malicious or unrestricted generative AI tools that can help create convincing phishing emails and deceptive websites (p. 5). Supports the "AI as deception multiplier" argument.

Erukude, Marella, & Veluru. (2026). AI-Driven Cybersecurity Threats: A Survey of Emerging Risks and Defensive Strategies. Academic survey paper. AI-enabled threats and defenses. Discusses AI-driven phishing, malware, ransomware, deepfakes, botnets, obfuscation, and defensive strategies (pp. 6–9). Useful as a broad academic literature-review source.

NIST. (2023). Artificial Intelligence Risk Management Framework 1.0. Government AI governance framework. AI risk governance and lifecycle management. Organizes AI risk management around four

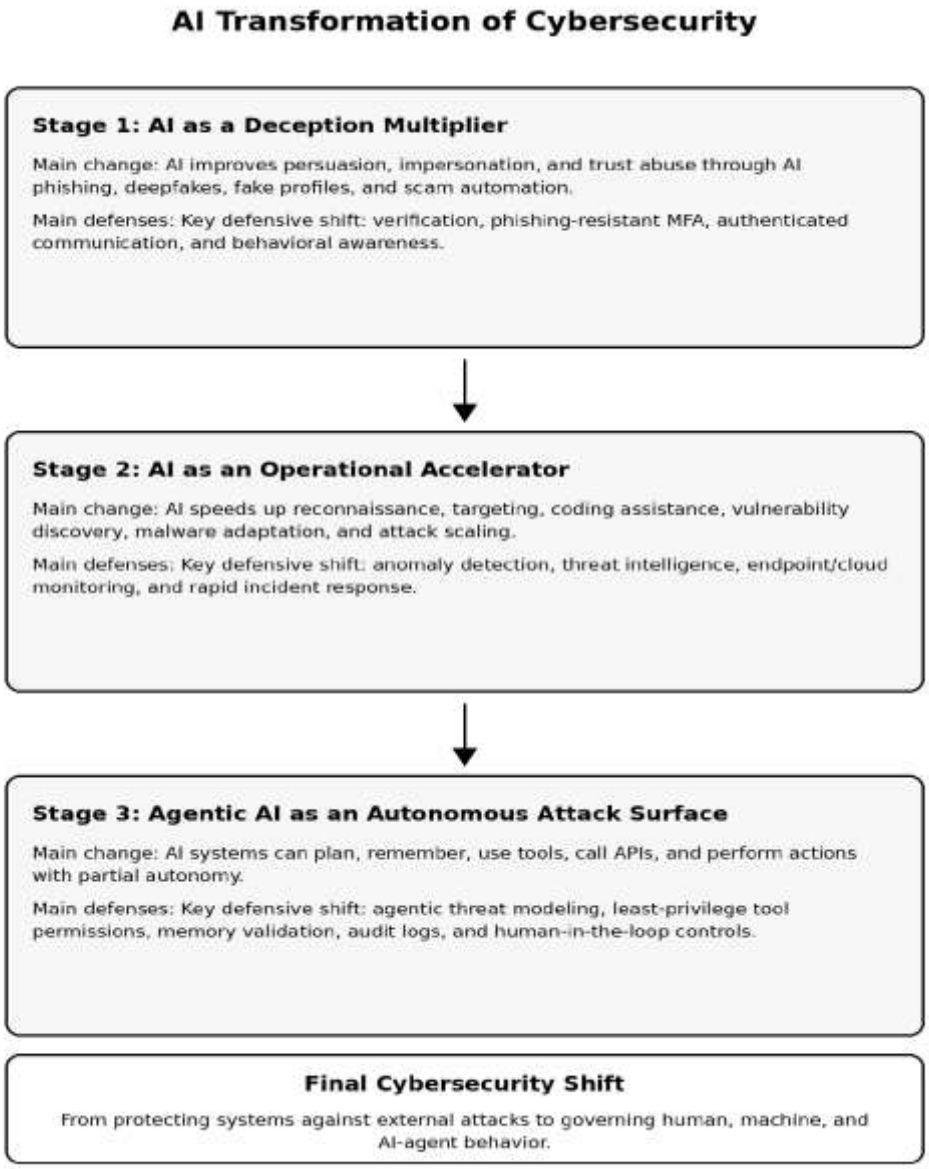
core functions: Govern, Map, Measure, and Manage (pp. 20–32). Useful for governance, risk management, and recommendation sections.

NCSC et al. (2023). Guidelines for Secure AI System Development. International government guidance. Secure AI lifecycle. Recommends secure design, secure development, secure deployment, and secure operation and maintenance across the AI system lifecycle (p. 8). Useful for defense and mitigation recommendations.

Chapter 4: Conceptual Framework

This paper proposes a three-stage conceptual framework to explain how AI is transforming cybersecurity. The framework moves from AI as a tool for deception to AI as a tool for operational acceleration to agentic AI as an autonomous attack surface.

Figure 1
Three-Stage Framework: From Phishing Bots to Autonomous Agents



AI as a Deception Multiplier

The first revolution in terms of changes brought about by AI concerns deception. In phishing, attackers have traditionally relied on mass messaging, template use, mistakes in spelling, questionable links, and impersonation. All of these strategies are still widely used, but generative AI technology enables malicious actors to compose more realistic-looking messages with lower costs and effort. With generative AI, one can mimic tone, translate from one language into another, adapt messages to specific roles, and create natural-looking communication as opposed to older phishing methods.

An academic paper examining the issue of ChatGPT, FraudGPT, and WormGPT used in social engineering attacks claims that generative AI technology can be used to personalize lures for social engineering attacks, influence opinions using deepfakes, exploit cognitive biases, and imitate authorities or organizations (Falade, 2023, p. 5). This is particularly alarming since social engineering does not relate to technology vulnerabilities but rather to psychological tricks of the trade. Social engineers rely on such human traits as trust, urgency, authority, fear, and curiosity.

There is no better proof for this claim than the report published by Microsoft. It argues that the advent of AI makes the problem of fraud and social engineering bigger since the technology is making the latter two faster and easier. Moreover, according to the report, the use of AI for the creation of synthetic identities, fake websites, fake customer service chat, and even deepfake fraud is one of the biggest problems in the current threat environment (Microsoft Threat Intelligence, 2025, p. 33). This does not mean that all AI-generated phishing attacks will be successful.

The threat of deepfakes and synthetic identities further complicates the issue. Synthetic voices, images, videos, profiles, websites, and customer interactions via chat pose even greater difficulties for users in separating manipulation from legitimate conversations. The 2025 Microsoft report emphasizes the possibility of using artificial intelligence in scammers' activity by creating fake websites, profiles, customer service conversations, as well as deepfake audio or video of trusted persons or companies (Microsoft Threat Intelligence, 2025, p. 33). ENISA, in turn, points out the employment of deepfakes, artificial intelligence personalities, as well as jailbroken large language models for social engineering and tool creation (ENISA, 2025, pp. 15-16).

This suggests that cybersecurity awareness can no longer focus only on obvious red flags such as spelling mistakes or suspicious email formatting. AI reduces those visible weaknesses. The new defensive question is not simply, “Does this message look fake?” but “Is this request behaviorally, contextually, and procedurally legitimate?”

From Phishing to AI-Assisted Cybercrime

The second transformation is operational acceleration. AI does not only help attackers communicate better; it can also support different stages of the attack lifecycle. These include reconnaissance, target profiling, malware assistance, vulnerability research, evasion, and post-compromise activity.

Mandiant’s M-Trends 2026 report is useful because it adds realism to the discussion. It does not claim that AI was the direct cause of most breaches in 2025. Instead, it says threat actors increasingly used AI tools for productivity gains, especially in reconnaissance, social engineering, and malware development (Mandiant, 2026, pp. 11–13). This is a more credible and balanced finding. AI may not yet be the sole driver of most successful intrusions, but it is becoming a force multiplier.

However, the Mandiant report reveals a shift in the methods used for first access. Exploits continued to be the most frequently used method of first infection in 2025, with 32% of cases being attributed to them, followed by 13% attributed to voice phishing and 6% to credential theft (Mandiant, 2026, p. 11). This is

important since it proves that the attackers are not only targeting the inbox anymore. Social engineering now involves interactions through voice, messages, job applications, browser prompts, and tech support. Similarly, ClickFix has been reported to be one of the key social engineering tactics in Microsoft's 2025 report. This tactic forces users to copy and execute some commands under the guise of solving a problem or verifying some information. According to Microsoft, ClickFix is a method where users are misled using false pop-ups or verification prompts to execute malicious commands, making phishing protection ineffective against this kind of manipulative behavior (Microsoft Threat Intelligence, 2025, p. 36). ENISA also reports the rise of ClickFix-type scams, which are usually hidden in fake CAPTCHA requests that force victims to execute PowerShell commands (ENISA, 2025, p. 16).

Such a development is crucial since most organizations focus on their defense systems against emails, links, and malware signatures. Techniques like ClickFix and vishing manipulate the user, who plays an active role in the process. Victims are coerced to carry out the harmful activity themselves. AI contributes to this trend by adding credibility to the commands given.

Ransomware further highlights how artificial intelligence operates within a wider criminal ecosystem. According to Mandiant, ransomware-related intrusions made up 13% of all cases Mandiant investigated in 2025, revealing extortion to be one of the most destructive types of cybercrime (Mandiant, 2026, p. 29). The finding revealed by Mandiant regarding a dramatic reduction in the time period between opportunistic initial access and hand-off to a different threat group is indicative of just how limited the window is for defense purposes (Mandiant, 2026, pp. 11-13).

Therefore, AI's second impact is not only better phishing. It is faster cybercrime. Attack chains become more scalable, more adaptive, and harder to interrupt before escalation.

Autonomous Agents as a New Attack Surface

The third and most novel transformation is the emergence of agentic AI. Unlike generative AI, which normally reacts to the prompt given to it, agentic AI is more advanced since it can plan tasks, utilize tools, connect to external services, search for information, retain memories, invoke APIs, and operate in digital environments. Such capabilities make agentic AI both helpful and dangerous.

According to OWASP's Agentic AI: Threats and Mitigation, agents have a number of capabilities, such as long-term and short-term memory, access to tools, APIs, databases, vector stores, web browsers, and multi-agent collaboration. This makes agents susceptible to security threats, as now there is no need for the attacker to hack into the underlying model. Instead, the hacker can trick the agent in various ways.

That is because agentic AI security involves additional threats that cannot happen with common chatbots. While a chatbot can misinform users with incorrect information, an agent can misinform users, send emails, collect data, change the record, activate workflows, and even reveal sensitive data. It moves the risks of misinformation from harmful content to harmful actions.

Some of the threats that OWASP mentions include memory poisoning, which refers to false/malicious information inserted into an agent's memory affecting its decision-making process; tool misuse when an attacker manages to make an agent use an authorized tool in an unintended way; privilege compromise, which means inheriting/misusing excessive privileges; cascading hallucinations, which mean false output distribution across multiple agents; identity spoofing/identity compromise that means impersonation of an agent/abusing agent identities; and lastly, human-in-the-loop overload when an attacker makes advantage of human limitations to produce numerous interactions (OWASP, 2025, pp. 23-41).

These risks reveal a major weakness in conventional security thinking. Traditional cybersecurity often focuses on securing users, endpoints, networks, applications, and data. Agentic AI adds another entity: a

semi-autonomous digital actor. This actor may have credentials, memory, goals, tools, and access rights. In that sense, an AI agent is not only software; it becomes an operational participant inside the organization.

CISA's AI Cybersecurity Collaboration Playbook reflects this concern by asking organizations reporting AI incidents to include details about affected AI artifacts, models, agentic platforms, copilots, third-party platforms, APIs, libraries, external systems, and the scope of information lost or exploited (CISA, 2025, p. 14). The playbook's prompt-injection example involving MathGPT further demonstrates how AI application vulnerabilities can create real operational disruption (CISA, 2025, p. 21). This shows that AI incidents require a broader evidence model than traditional incidents. Investigators may need to know not only what server was compromised, but also what model was used, what tools the agent could access, what data it retrieved, what memory it stored, and what actions it performed.

Agentic AI therefore marks a shift from cybersecurity as protection against external attackers to cybersecurity as governance over autonomous internal actors.

AI as a Defense Multiplier

While it enhances attackers, AI also enhances defenders. Cybersecurity teams are confronted with huge amounts of alerts, logs, vulnerabilities, identities, endpoints, cloud assets, and threat intelligence. A human analyst would be incapable of processing such information at modern speeds. AI can assist with anomaly detection, prioritization of alerts, summarization of logs, fraud detection, malware classification, correlation, and response acceleration.

"Digital Defense Report 2025" by Microsoft claims that AI allows for synthesis of enormous datasets, detection of new threats, and acceleration of response (Microsoft Threat Intelligence, 2025, pp. 52-53). Mandiant underlines the need for defenders to move past static defenses and improve monitoring of identity behavior, infrastructure, and environments that cannot be covered by traditional endpoint tools (Mandiant, 2026, pp. 11-13).

Nonetheless, there is an inherent limitation to using AI for defense. Automated defenses might create false alarms, ignore new types of behavior, generate fake explanations, and even be overly trusted by their human counterparts. Models of artificial intelligence are vulnerable to attacks through prompt injection, data poisoning, adversarial examples, model extraction, and supply chain attacks. In the Generative AI Profile published by the National Institute of Standards and Technology, it is highlighted that "generative AI reduces the barrier to offensive cyber operations by increasing the attack surface through prompt injection, data poisoning, malware, hacking, and phishing" (NIST, 2024, p. 10).

The OWASP Top 10 for LLM Applications 2025 identifies additional risks associated with prompt injection, sensitive data leakage, supply chain risk, data and model poisoning, output handling failure, too much agency, vectors and embedding flaws, misinformation, and unbounded consumption (OWASP, 2024, pp. 3–5, 22–28). Clearly, such issues prove why an organization cannot just incorporate AI into security operations without proper governance, since uncontrolled AI-based defense mechanisms can leak data, misclassify attacks, or be manipulated through adversarial input.

This means that the real question is not whether AI favors attacks or defense operations. The correct answer is that AI makes things move faster on both sides, and the winning party will be the one that wins in terms of governance, identity control, telemetry, response discipline, and human-machine cooperation.

Governance and Risk Mitigation

The available frameworks suggest that organizations need a layered response.

First, organizations must secure identity. Microsoft's report emphasizes phishing-resistant multifactor au-

thentication and identity protection as key controls (Microsoft Threat Intelligence, 2025, pp. 52–53). This matters because many modern intrusions do not begin with attackers "breaking in" through technical exploits; they begin with attackers logging in using stolen credentials, tokens, or abused identities.

Thirdly, there is a need to enhance behavioral detection capabilities in organizations. Traditional static signals cannot work well in situations where threats leverage legitimate mechanisms, social engineering tactics, or even user commands. Threats such as ClickFix, voice phishing, and living off the land demand that security measures should focus on detecting anomalous behavior.

Fourthly, there is a need to secure AI systems across the lifecycle. The guidelines for secure AI system development by the National Cyber Security Centre indicate the importance of secure design, secure development, secure deployment, and secure operation and maintenance of AI systems (NCSC et al., 2023, p. 8). This is because potential threats can emerge at different stages of the lifecycle.

Fourth, organizations must adopt AI risk management frameworks. NIST's AI Risk Management Framework organizes AI risk management around four functions: Govern, Map, Measure, and Manage (NIST, 2023, pp. 20–32). These functions help organizations identify AI risks, assess their impact, measure system behavior, and implement appropriate controls.

Fifth, organizations must threat model agentic AI specifically. Agentic systems require controls over memory, tool use, permissions, identity, logging, inter-agent communication, and human oversight. OWASP's agentic AI guidance recommends memory validation, session isolation, strict tool access verification, rate limiting, audit logs, granular permissions, behavioral monitoring, and constraints on inter-agent delegation (OWASP, 2025, pp. 23–41).

The sixth step organizations should take concerns improving incident reporting and cooperation. The AI playbook issued by CISA stresses the importance of voluntary information exchange, vulnerability coordination, incident reporting, and inclusion of specific AI artifacts within incident reports (CISA, 2025, p. 14). Such an approach is crucial since AI threats are fast-evolving and usually transcend organizational borders.

The last change needed concerns rethinking human oversight. Human-in-the-loop does not presuppose that a person will be required to agree on numerous complex decisions without any context. Otherwise, such a decision may be exploited by the attacker.

Chapter 5: Recommendations

Level	Recommended Control	Why It Matters	Related Source Base
Individual	Use phishing-resistant MFA wherever possible.	AI makes phishing and impersonation more convincing, so password-only security is insufficient.	Microsoft, Mandiant
Individual	Verify unusual requests through a second trusted channel.	AI-generated messages, deepfakes, and fake profiles can imitate trusted people or institutions.	Microsoft, ENISA, Falade

Individual	Avoid copying commands from pop-ups, CAPTCHA prompts, or unknown support messages.	ClickFix-style attacks exploit users by persuading them to execute malicious commands themselves.	Microsoft, ENISA
Individual	Treat voice, video, and image-based verification as fallible.	Deepfakes and synthetic media weaken traditional trust signals.	Microsoft, ENISA, NIST GenAI Profile
Individual	Develop behavioral awareness rather than relying only on spelling or formatting errors.	AI-generated phishing can remove many obvious signs of fraud.	Microsoft, Falade
Organizational	Implement phishing-resistant MFA and strong identity governance.	Many attacks begin through stolen credentials, abused identities, or social engineering rather than purely technical exploits.	Microsoft, Mandiant
Organizational	Move from signature-based detection to behavior-based monitoring.	AI-assisted attacks and ClickFix-style techniques may bypass static indicators.	Microsoft, Mandiant
Organizational	Maintain strong endpoint, cloud, SaaS, and identity telemetry.	Modern attacks often move across cloud accounts, SaaS systems, endpoints, and third-party platforms.	Microsoft, Mandiant
Organizational	Conduct AI-specific threat modeling before deploying LLM or agentic systems.	Agentic AI introduces risks involving memory, tools, permissions, APIs, and autonomous workflows.	OWASP Agentic AI, CISA
Organizational	Restrict AI-agent tool permissions using least privilege.	Agents should not have broad access to data, email, APIs, files, or administrative tools unless required.	OWASP Agentic AI, OWASP LLM Top 10
Organizational	Validate and monitor agent memory.	Persistent memory can be poisoned and may influence future agent behavior.	OWASP Agentic AI
Organizational	Keep audit logs of AI-agent decisions, tool calls, and data access.	Agentic systems must be traceable because they can perform actions	OWASP Agentic AI, CISA

		with real operational consequences.	
Organizational	Use human-in-the-loop controls for high-impact AI actions.	Human oversight is necessary for risky actions, but it must be designed to avoid decision fatigue.	OWASP Agentic AI
Organizational	Follow secure AI lifecycle practices.	AI systems should be secured during design, development, deployment, operation, and maintenance.	NCSC, NIST
Policy / Governance	Adopt AI risk management frameworks.	Organizations need a structured way to govern, map, measure, and manage AI risks.	NIST AI RMF
Policy / Governance	Standardize AI incident reporting.	AI incidents require details about models, datasets, APIs, tools, platforms, and access rights.	CISA AI Playbook
Policy / Governance	Encourage public-private threat intelligence sharing.	AI-related cyber threats evolve quickly and require collaboration across government, industry, and researchers.	CISA, ENISA
Policy / Governance	Create sector-specific guidance for critical infrastructure.	AI risks in healthcare, finance, transport, energy, and public services can have physical and social consequences.	NCSC, NIST, academic literature
Policy / Governance	Require transparency and documentation for high-risk AI systems.	Documentation helps organizations understand AI behavior, dependencies, data exposure, and security responsibilities.	NIST AI RMF
Policy / Governance	Promote responsible red teaming and adversarial testing of AI systems.	AI systems should be tested for prompt injection, data leakage, tool misuse, and adversarial manipulation before deployment.	OWASP, NIST, CISA

Chapter 6: Discussion

The evidence indicates that artificial intelligence will not create an entirely new cyber world all at once. What it does is amplify current trends while introducing new categories of risk. Phishing attacks are still important, but with AI, they are made more effective. Ransomware attacks are still significant, but with

AI, they may be sped up through reconnaissance and social engineering. Identity is important too, but autonomous agents present identities outside humans.

One good way to think about this transformation is by considering three levels of operation. First, the AI acts as a deception mechanism by enhancing phishing, deepfakes, identity creation, impersonation, and personalized scams. Second, AI acts as an operational mechanism by providing support for reconnaissance, malware, vulnerability studies, evasion, and scalable cybercrimes. Third, AI acts as an attack surface autonomously by creating risks associated with memory, tools, permissions, identity, and autonomous actions.

Using the three-level approach makes it easier to sidestep two unconvincing arguments. The first one is technological panic, where people assume that AI has rendered all existing cybersecurity solutions obsolete. The second one is technological optimism, where people believe that AI will solve all cyber threats simply because it can detect all attacks. Neither argument is supported by the evidence presented so far.

The most important shift is from content security to action security. Traditional phishing defense asks whether a message is malicious. Agentic AI security must ask whether an autonomous system should be allowed to perform a particular action, with a particular tool, using particular data, under particular conditions. This is a deeper governance problem.

Chapter 7: Conclusion

AI is transforming cybersecurity from both sides of the conflict. Attackers use AI to improve deception, accelerate operations, and exploit human trust. Defenders use AI to process data, detect anomalies, prioritize risk, and respond faster. The result is not simply an “AI versus human” conflict but an increasingly AI-mediated security environment in which both attackers and defenders use automation.

“From Phishing Bots to Autonomous Agents” is the theme of this progression. First comes deception by means of AI. Then comes cybercrime facilitated by AI. Finally, there is risk through autonomous agency. As AI systems develop more memory, more tools, more permissions, and the capacity for action, cybersecurity must shift from safeguarding networks and individuals to governing semi-autonomous digital agents.

It would be folly to rely on AI alone to fend off any cyber attack. In addition to phishing-proof identity controls, detection based on behavior, safe AI development practices, threat modeling for AI, agency permission management, memory integrity management, audit logging, human oversight, and incident reporting coordination will all play a critical part in creating a cybersecure future.

AI might expedite and amplify cyberwarfare, but fundamentally the problem is human one and thus requires human solutions.

The facts clearly indicate that AI is not substituting for cybersecurity but changing the environment within which it takes place. Traditional cyber attacks like phishing, credential attacks, ransomware, malware, and social engineering remain highly pertinent. It is only the pace, scale, customization, and adaptation of these attacks that have undergone modification. Generative AI enhances deception through more effective creation of scams, false personas, deepfakes, and phishing communications. AI also helps speed up the process of cyberattacks through reconnaissance, vulnerability identification, malware creation, evasion, and increased productivity of attackers. The most critical new trend is agentic AI in which machines can use tools, store information, call API, assign tasks, and act autonomously.

Three elements within the framework used in this paper – AI as a deception multiplier, AI as an operational accelerator, and agentic AI as an autonomous attack surface – explain the evolution from phishing bots to autonomous agents. This framework explains why cybersecurity needs to change. Email filtering and malware signatures continue to be relevant in defending against threats. However, they are no longer enough. What is needed now is identity-based cybersecurity, behavioral detection, secure development practices for AI, specific threat modeling for AI, agent permissions management, memory validation, logging and auditing, human in the loop governance, and incident reporting coordination.

Cybersecurity in the future will not be about AI. Cybersecurity in the future will be about governing AI-enabled risks. AI can enhance attackers, and it can also enhance defenders. Everything will depend on how the risk will be governed and made visible.

References

1. Cybersecurity and Infrastructure Security Agency. (2025). JCDC AI Cybersecurity Collaboration Playbook. U.S. Department of Homeland Security.
2. European Union Agency for Cybersecurity. (2025). ENISA Threat Landscape 2025.
3. Erukude, S. T., Marella, V. C., & Veluru, S. R. (2026). AI-driven cybersecurity threats: A survey of emerging risks and defensive strategies.
4. Falade, P. V. (2023). Decoding the threat landscape: ChatGPT, FraudGPT, and WormGPT in social engineering attacks. International Journal of Scientific Research in Computer Science, Engineering and Information Technology.
5. Lohn, A. J. (2025). The impact of AI on the cyber offense-defense balance and the character of cyber conflict. Center for Security and Emerging Technology, Georgetown University.
6. Mandiant. (2026). M-Trends 2026 Special Report. Google Cloud Security.
7. Microsoft Threat Intelligence. (2025). Microsoft Digital Defense Report 2025: Safeguarding trust in the AI era. Microsoft.
8. National Cyber Security Centre, Cybersecurity and Infrastructure Security Agency, National Security Agency, Federal Bureau of Investigation, Australian Cyber Security Centre, Canadian Centre for Cyber Security, National Cyber Security Centre New Zealand, & international partners. (2023). Guidelines for secure AI system development.
9. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce.
10. National Institute of Standards and Technology. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). U.S. Department of Commerce.
11. Open Worldwide Application Security Project. (2024). OWASP Top 10 for LLM applications 2025.
12. Open Worldwide Application Security Project. (2025). Agentic AI: Threats and mitigations.
13. Open Worldwide Application Security Project. (2026). State of agentic AI security and governance.
14. Paulraj, J., Raghuraman, B., Gopalakrishnan, N., & Otoum, Y. (2025). Autonomous AI-based cybersecurity framework for critical infrastructure: Real-time threat mitigation.