International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 10 | Oct 2026 www.irjet.net p-ISSN: 2395-0072

AI vs Agentic AI: The Future of Autonomous Cyber-Security

Mr. Yash R. Mali[1], Ms. Priyanka M. Khalkho[2], Prof. Mrs. Swati Kadu[3], Mr. Nimesh R. Apte[4]

[1234]DY Patil University School of Humanities and Sciences, Navi Mumbai, India

---

Abstract - *Artificial intelligence now shapes both sides of cybersecurity. Agentic AI, meaning systems that set sub-goals, choose actions and carry them out with little human involvement, extends this change further and brings both new capability and new risk. This paper traces how AI in security developed, from rule-based expert systems to agent-based architectures, and compares how it is applied in cyber offense and cyber defense. It then considers what conflict between offensive and defensive agents may look like. The review suggests that agentic systems can improve detection speed and response time, but that they also add attack surfaces of their own, such as poisoned memory and hijacked tool calls. Governance, human oversight and protection of the defending AI are therefore as important as raw capability.*

Key Words: Cyber Attack, Cyber Defense, Cyber Security, Security Operations, Cyber Warfare, Artificial Intelligence (AI), Agentic AI.

Objectives: To study the Evolution of AI in Cyber Security. To Analyse the use of AI in Cyber Offense and Cyber Defence. To Examine the concept of AI vs Agentic AI in future Cyber warfare.

1. INTRODUCTION

Threats in cyberspace keep growing in scale and sophistication, and attackers continually refine their methods. Traditional AI-based defenses, which mostly perform prediction and pattern recognition, have strengthened security operations considerably, yet they still depend on human analysts for significant decisions [3].

Agentic AI changes this arrangement. An agentic system can observe its environment, choose a course of action and execute it without waiting for a person [2]. In security work this allows real-time anomaly detection, correlation across large data sets and fast containment of threats. The same autonomy raises ethical and security questions, because an agent that can act on its own can also be misled or misused [4].

2. EVOLUTION OF AI IN CYBERSECURITY

2.1 Traditional Rule-Based Systems

The idea of using statistical profiles of normal behavior to spot misuse goes back to Anderson's 1980 report on computer security threat monitoring for the US Air Force [7]. Denning's 1987 intrusion-detection model formalized this approach, and SRI International implemented it in the Intrusion Detection Expert System (IDES) [8]. Rule-based expert systems of this period encoded the knowledge of human specialists as "if-then" rules held in static knowledge bases [1].

This design had clear weaknesses. Such systems struggled with unfamiliar attack patterns, needed constant manual updating of their rules (the knowledge acquisition bottleneck), and produced many false alarms [1]. These limits, together with wider disappointment in the field, coincided with the second AI winter of the 1990s [1], [2].

2.2 The Statistical and Machine Learning Era (Late 1990s–2010s)

Growing connectivity and data volumes shifted research toward learning from data. Public benchmark datasets, notably those released by DARPA in 1998 and 1999, gave researchers common ground for testing supervised and unsupervised detection methods [2].

- **Supervised learning.** Supervised learning used labeled examples for spam filtering, phishing detection and later antivirus engines, with models such as decision trees, support vector machines, random forests and gradient boosting [2], [3].
- **Unsupervised learning.** Unsupervised learning looked for outliers in otherwise normal traffic in order to flag unknown threats, but was limited by high false-positive rates and the alert fatigue they cause among analysts [3].
- **Behavioral analytics.** By the mid-2010s, user and entity behavior analytics (UEBA) was being added to SIEM platforms to build baselines of normal activity and expose stolen credentials or insider misuse [2].
- **Deep learning.** Neural architectures then entered defensive tooling: convolutional networks that treat malware binaries as images, recurrent networks and LSTMs for sequences such as logs, and generative adversarial networks for producing synthetic training data [3].
- **Adversarial machine learning.** As defenders adopted ML, attackers began targeting the models themselves through data poisoning, evasion with crafted inputs, and model extraction [3], [5].

2.3 The Generative AI Revolution (2022–2024)

The transformer architecture and models such as BERT made large language models (LLMs) practical, and the public release of ChatGPT in November 2022 accelerated adoption by defenders and attackers alike [6]. Defensively, LLMs have been applied in security operations centers for natural-language log queries, summarizing alerts, drafting incident-response plans, reviewing code and powering adaptive honeypots [2], [6].

Attackers have used the same technology for deepfake and voice-cloning fraud, more convincing phishing, and malware that changes its own form [6]. LLM-based systems also introduced their own weaknesses, including prompt injection, indirect prompt injection and the hijacking of tool calls [6].

2.4 Agentic AI and Modern Attack Surfaces (2025–Present)

The present stage moves from systems that analyze text passively to goal-directed, often multi-agent systems built on LLMs. These can run multi-step investigations, call external tools and carry out threat hunting on their own [2], [4]. This autonomy has produced new classes of attack:

- **Agent memory poisoning.** An attacker plants instructions in content that an agent later stores in long-term memory, so the compromise takes effect in a later session. This is catalogued in OWASP's guidance on agentic applications [9].
- **Coding-agent hijacking.** Malicious text placed in data that a developer's coding assistant reads, such as logs or error output, can persuade the assistant to run attacker-chosen commands with the user's privileges [9].
- **Cross-agent session smuggling.** Commands can be passed across trust boundaries between cooperating agents without authorization [9].

Several technical and governance measures respond to these risks. Software-defined networking and network function virtualization allow detection models to be deployed and updated centrally in programmable network controllers [3]. Organizations are also adopting the adversarial machine learning taxonomy in NIST AI 100-2 [10], explicit identity management for agents using standards such as OAuth 2.0 and SPIFFE/SPIRE, container sandboxing, and mandatory human-in-the-loop approval for high-impact actions [9], [10].

3. ANALYZE THE USAGE OF AI IN CYBER OFFENSE AND CYBER DEFENCE

AI brings clear benefits to security, including faster detection and blocking of threats, continuous monitoring that does not tire, quicker identification of unusual activity, analysis of data volumes beyond human capacity, and fewer human errors [3]. It also enables new forms of attack. This section considers both sides.

3.1 The Use of AI in Cyber Offense

Offensive uses fall into two groups: assisting human attackers and carrying out attacks autonomously [5].

3.1.1 Helping Human Attackers

Here AI increases the speed, accuracy and reach of a human operator without replacing them. It can process social media, websites and public records to build profiles of targets, which supports more believable phishing and social engineering [5], [6]. It can help write or modify malicious code, and it can automate vulnerability scanning so that weaknesses are found and exploited with little manual effort [5], [6].

3.1.2 Performing Attacks Automatically

Agentic AI goes a step further by carrying out intrusion steps itself rather than advising an operator. Given tools and a goal, an agent can identify targets, plan the sequence of actions and execute them [4]. Its main advantages are speed, round-the-clock operation, the ability to attack many systems at once and the ability to adapt as conditions change. The risk is greatest where unpatched systems exist, because automated discovery combined with automated exploitation shortens the time victims have to respond [4], [5].

3.2 The Use of AI in Cyber Defense

Traditional security tooling often cannot keep pace with the volume and complexity of modern attacks. AI helps organizations detect, prevent and respond more quickly [3].

3.2.1 Threat Detection and Monitoring

Continuous monitoring of networks, devices and user behavior is the most widespread defensive use. A login from an unfamiliar location or device, for example, can be flagged for review immediately [3]. AI-based intrusion detection systems analyze traffic, system activity and user behavior in real time and can detect threats more accurately than purely rule-based systems [2], [3].

3.2.2 Automated Response

Manual remediation can take hours, which gives an attacker time to move through a network. Automated response offers speed and consistency. A common design grades each threat by severity and selects an action accordingly [4], as summarized in Table -1.

Table -1: Automated response actions by risk level

Risk level	Typical automated action
Low	Continue monitoring and analysis
Medium	Soft containment, alerting and flagging for review
High	Immediate blocking and disabling of the affected account or system
Critical	Human approval or strict policy checks before any action

3.2.3 SOC Operations and Human Factors

AI helps security teams by triaging alerts, cutting false alarms, ranking threats by priority and supporting analysts. It can also be used to train employees to recognize phishing, and analysts are more likely to trust its output when its decisions can be explained [3], [4]. Combining machine speed with human judgment makes security operations more effective.

3.2.4 Securing the AI Defenders Themselves

An AI defender is itself a target. If attackers can deceive, manipulate or damage it, the security layer becomes a weakness [5], [10]. Defending AI systems therefore needs the same attention as any other critical asset.

Attackers can use AI to find vulnerabilities, build attacks and execute them at machine speed. Defenders use it to detect threats, analyze large volumes of data and respond quickly. Because human teams cannot match that speed across everything, security is increasingly a contest in which AI systems defend against other AI systems [4]. The main challenge is therefore not making AI more powerful but making autonomous AI trustworthy and safe, which calls for rigorous evaluation, governance, monitoring, human oversight and protection of the defending systems themselves [4], [10].

4. AI VS AGENTIC AI IN FUTURE CYBER WARFARE

Future cyber conflict may look less like human hackers against static code and more like a continuous, fast exchange between offensive and defensive agents [4].

4.1 The Digital Arena

The contested space covers endpoints, networks, cloud platforms, user identities and data channels [4].

4.2 How Attackers Move (Offensive AI)

Reported offensive capabilities include decision-making driven by reinforcement learning, automated scanning with rapid exploit delivery, payloads that change form to evade filters, covert command channels, and continuous real-time adaptation [4], [5].

4.3 How Defenders Fight Back (Defensive AI)

Defensive capabilities include round-the-clock vigilance, proactive threat hunting, flexible choice of response, deception through decoy networks and misdirection, and predictive risk scoring [3], [4].

4.4 The Adaptive Loop

Both sides improve through a repeated cycle: observe the situation, predict the opponent's intent, act (attack or counter) and evaluate the result so that strategy can be adjusted [4]. Over time this makes security a self-learning ecosystem in which each side's progress pushes the other to improve.

4.5 Training AI Security Agents

Agents need realistic practice before facing real conflict. Offensive research agents are commonly trained in controlled settings such as Microsoft's CyberBattleSim, capture-the-flag exercises and reinforcement-learning sandboxes, which allow techniques to be practiced without risk to live systems [4]. Defensive agents learn from simulated attacks, past breach data and live network telemetry, and adversarial reinforcement learning helps them adapt when new attack patterns appear [4]. Two simple measures of effectiveness are detection rate, meaning how well threats are found, and response latency, meaning how quickly they are contained [4].

4.6 What Lies Ahead

The move toward agent-versus-agent conflict means that reactive defenses alone will not be enough. Several priorities follow from the literature [4], [3]:

- Networks of cooperating defensive agents that can protect hybrid environments spanning cloud, on-premises, IoT and edge systems.
- Transparent architectures that let analysts inspect and audit autonomous decisions.

- Bio-inspired defenses, drawing on immune systems and swarm behavior, that can repair themselves and adapt.
- Preparation for the quantum era, including post-quantum cryptography.
- Continued human-agent collaboration: machines contribute speed, while people contribute context, ethics and strategic judgment.

AI will not be a single solution to every security problem, at least in the near term. The most practical systems today rely on supervised learning and unsupervised anomaly detection, and the latter still needs expert analysts to interpret what an anomaly means [3]. The next generation is expected to be context-aware, using approaches such as Bayesian belief networks that consider how a system behaves in specific circumstances instead of only flagging outliers, supported by visual analytics that help analysts understand the data [3].

4.7 Offensive Applications Across the Cyberattack Lifecycle

Adversaries can apply AI at each stage of the cyber kill chain: automating reconnaissance, building evasion mechanisms that adapt, and speeding up execution [5]. In the early stages, learning-based methods can analyze telemetry to assess an environment and locate weaknesses. Research has also shown that models can be attacked directly, for example through membership inference, where an adversary probes a model to learn whether particular records were in its training data [5].

Generative AI lowers the skill needed for automated attacks, tailored phishing and social engineering [6]. Deepfakes can be used to influence opinion or defeat authentication, and generative models can help produce malware that changes its form [6]. DeepLocker, a proof of concept built by IBM Research, showed how malicious code could stay concealed inside an ordinary application until a specific trigger, such as a recognized face or location, released it [11]. In cyber-physical settings, manipulating inputs such as traffic signals or medical images could compromise safety without triggering conventional security controls [5].

5. WEAPONIZATION OF AI

5.1 The Weaponization of AI

AI is already being built into military systems, where it supports decision-making for platforms on land, at sea, in the air and in space [5]. The same literature links it to nuclear, chemical and toxic-materials research, next to peaceful uses such as climate control and space exploration [5]. Misuse in cyberspace is not hypothetical. Microsoft's Tay chatbot began producing biased and offensive replies after users fed it hostile input, and the 2016 US election drew attention to AI-driven manipulation of public opinion on social media [5]. A model trained on deliberately corrupted data, or one coupled to high-consequence weapons, could cause severe harm, which is why controls on its strategic deployment are needed [5].

In this context, weaponized AI can be defined as harmful algorithms built to undermine beneficial AI systems and to give their operators a strategic edge in digital or physical domains [5]. Cyberspace produces more data than people can watch, so AI is the only practical way to keep security within useful time frames, yet flaws in AI software and hardware carry serious consequences [5]. UC Berkeley researchers, for instance, showed that small changes to road signs could make an autonomous vehicle's vision system misread a stop sign [5]. A military vehicle relying on a similarly fragile model could have its view of the road distorted by an adversary. IBM Research's DeepLocker illustrates a different risk. It was a proof of concept that hid the WannaCry ransomware inside a video-conferencing application and released it only when a chosen target was recognized [11]. It was a demonstration of evasive AI-assisted malware, not a tool used in the real WannaCry outbreak.

5.2 Weaponized AI in Cyberspace

The literature distinguishes two ways of weaponizing AI in cyberspace [5]. One embeds it in established military doctrine. The other deploys it during operations alongside quantum computing, big data and robotics. The second route is regarded as the more realistic one today, because such combinations allow analysis in real time, faster battlefield decisions, steadier logistics and the delegation of hazardous work, and together they could shift the balance of power in conventional operations [5].

The same source argues that the idea of "mass" changes in this setting: the volume of analysable data, more than troops or equipment, becomes the measure of strength, because models tend to improve as data grows [5]. As weaponization advances, the armed services will lean on AI more heavily. People are expected to keep the key deployment decisions for now, but more control is likely to pass to machines over time, which makes regulation and oversight necessary [5]. For that reason, an explicit AI strategy and study of the consequences of weaponized AI are recommended so that suitable usage controls can be designed [5].

5.3 AI-Powered vs. Classic Cyberattacks

Classic attacks include denial of service (DoS and DDoS), man-in-the-middle attacks, phishing and spear-phishing, injection attacks such as SQL injection and cross-site

scripting, jamming, eavesdropping and malware. They are well documented and fall outside the scope of this paper. The focus here is on attacks powered by AI, which Yamin et al. [5] sort into three groups: data misclassification, synthetic data generation and data analysis. Because AI depends on data, the scenarios discussed involve audio, visual and textual inputs, although the techniques generalize to other data. The STRIDE threat model [12] provides a common basis for comparing AI-powered attacks with classic ones.

6. CONCLUSION

AI has moved from rule-based expert systems to learning systems, then to generative models and now to agents that act with growing independence. Each stage improved defense and gave attackers new tools. As both sides adopt autonomous agents, security may increasingly become a contest between AI systems operating at machine speed. The central challenge is to make such systems trustworthy, which requires strong evaluation, transparent design, governance, protection of the defenders themselves and continued human oversight.

ACKNOWLEDGEMENT

The authors sincerely thank the Department of MCA, D Y Patil University School of Humanities and Sciences, for the academic support, facilities, and research environment provided during this work.

REFERENCES

1) A. Quffa and S. S. Abu-Naser, "A rule-based expert system for cybersecurity threat detection: Evolution, applications, and the hybrid AI paradigm," Int. J. Academic Engineering Research (IJAER), vol. 9, no. 8, pp. 44-62, Aug. 2025.

2) V. Anand, "From expert systems to agentic AI: The evolution of AI in cybersecurity," Splunk Blog, Feb. 25, 2026. [Online]. Available: https://www.splunk.com/en_us/blog/artificial-intelligence/evolution-of-ai-in-cybersecurity.html

3) L. Jain, "Cybersecurity: Transitioning from conventional defense to intelligent threat management," Int. Research J. of Engineering and Technology (IRJET), vol. 12, no. 10, Oct. 2025.

4) S. A. Mohammed, "Agent-vs-agent cyber warfare: Autonomous AI systems defending against AI-enabled APTs," Int. J. on Advanced Computer Engineering and Communication Technology, 2025.

5) M. M. Yamin, M. Ullah, H. Ullah, and B. Katt, "Weaponized AI for cyber attacks," J. Information Security and Applications, vol. 57, 2021.

6) M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, "From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy," IEEE Access, vol. 11, 2023.

7) J. P. Anderson, "Computer security threat monitoring and surveillance," James P. Anderson Co., Fort Washington, PA, USA, Tech. Rep., 1980.

8) D. E. Denning, "An intrusion-detection model," IEEE Trans. Software Engineering, vol. SE-13, no. 2, pp. 222-232, Feb. 1987.

9) OWASP GenAI Security Project, "OWASP Top 10 for agentic applications," 2025/2026. [Online]. Available: https://genai.owasp.org

10) A. Vassilev et al., "Adversarial machine learning: A taxonomy and terminology of attacks and mitigations," NIST AI 100-2 E2025, 2025.

11) D. Kirat, J. Jang, and M. Stoecklin, "DeepLocker: Concealing targeted attacks with AI locksmithing," presented at Black Hat USA, 2018.

12) A. Shostack, Threat Modeling: Designing for Security. Indianapolis, IN, USA: Wiley, 2014.