CENTRE FOR AEROSPACE POWER AND STRATEGIC STUDIES

In Focus

New Delhi

CAPSS In Focus: 43/2026

04 September 2026

Agentic AI and Cyber Espionage: Incidents, Implications & Recommendations

Ms Gowri R
Research Associate, Centre for Aerospace Power and Strategic Studies

Source: Created by the Author using AI

Keywords: AI-Powered Cyberattacks, Agentic AI, Cyber Incidents, Cyber Defence

Introduction

The rapid advancement of autonomous Artificial Intelligence (AI) systems has introduced a new frontier in cyber espionage, where intelligent machines can independently plan and execute sophisticated attacks. Defence Information and Communication Technology (ICT) infrastructure, including networks and systems, is a prime target for state and non-state attackers because it supports critical military functions such as readiness, decision-making, space operations, and Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C4ISR) architectures.[1] An intrusion into such systems may not immediately cause physical destruction, but it can create serious strategic effects by stealing classified information, corrupting or spoofing data, disrupting or denying services, delaying mobilisation, degrading situational awareness and forcing commanders to act on false information. This paper focuses on recent AI-powered cyber incidents reported by OpenAI and Anthropic, identifies implications and provides recommendations to improve the resilience of defence systems.

The New Phenomena

Imagine malware that doesn't just follow a script, but thinks on its feet. Traditional malware is like burglars who know one way to break into a house. It follows prewritten rules or code to infect, disrupt, spy on, steal from, or control a computer system in a known software or environment.[2] AI-powered malware, on the other hand, is like a master thief who studies your locks, adapts to your alarms, and even learns new tricks mid-break-in. It rapidly identifies vulnerabilities, processes target information, modifies tactics in response to defensive controls and coordinates multiple technical tasks with or without human intervention.[3] This is the world of AI-powered cyberattacks.

Agentic AI are autonomous AI systems that take these tactics to the next level. They can break down complex instructions into simpler ones and complete multiple tasks automatically, with minimal human intervention.[4] It also changes its tactics according to the changing environment and feedback.[5] The dual-use AI technology has the capability to autonomously discover and patch software defects and can also be directed to identify exploitable weaknesses.[6] Research by the United States (US) Congressional Research Service observes that AI agents lower the operational burden for sophisticated state actors.[7] It potentially expands the capability of less-resourced groups. The following are some notable agentic AI-powered cyber incidents reported and documented by leading AI industry organisations.

1. **China-linked GTG-1002 campaign (2025):**[8] In November, 2025, Anthropic, a US-based company, disclosed the cyber espionage campaign attributed to the GTG-1002 (Global Threat Group), a Chinese-sponsored group, as per Anthropic's assessment. This group reportedly used the Agentic AI capabilities of Claude Code to interpret complex instructions, autonomously execute and sequence multiple tasks and use cyber-related tools. It manipulated Claude Code and attempted intrusion against approximately 30 global targets, including technology companies, financial institutions, chemical manufacturers and government agencies. This cyber espionage has exploited agentic AI capabilities such as the ability to intercept complex instructions, autonomous execution and sequencing of multiple tasks and accessing cybersecurity-related tools. Anthropic estimated that AI performed 80 to 90 per cent of the campaign's work, while human involvement was at a small number of critical decision points. However, it also acknowledged limitations such as Claude sometimes hallucinating credentials or claiming to obtain sensitive information that was actually publicly available.

2. **ScopeCreep (2025):** OpenAI's *Disrupting Malicious Uses of AI June* 2025 report [9] describes cyber espionage as the use of AI to support the unauthorised collection of sensitive information from targeted networks, organisations or individuals. In the operation termed 'ScopeCreep', Russian-language actors reportedly used AI to assist in the development of a stealthy Go-based malware tool disguised as a gaming-related application.[10] This case illustrates how AI-powered malware development, coding assistance and operational deception can enable threat actors to conduct cyber operations more efficiently.

3. **Claude AI sandbox breach (2026):** In July 2026, Anthropic reported three incidents[11] after reviewing 141,006 evaluation runs, in which Claude models accessed the live internet from a third-party testing environment and gained unauthorised access to the production systems of three external organisations. Anthropic stated that the models treated real internet systems as elements of the simulated exercise. It exploited a real-world organisation that had been anonymised and presented under a fictional name. This case reveals that the AI agent, when connected to real-world tools and inadequately controlled infrastructure, can affect external systems beyond its intended environment.

Implications of Agentic AI-Powered Cyber Incidents

Today, AI systems can facilitate large-scale cyberattacks by automating surveillance, vulnerability discovery, and other malicious activities with limited human intervention. AI agents persist through many attempts to exploit basic weaknesses such as weak passwords, unauthenticated endpoints, exposed debugging information and SQL (Structured Query Language) injection. The increasing

accessibility of these capabilities enables a wider range of actors to target and exploit organisational vulnerabilities for inimical use.

Claude AI has provided hallucinated content containing user credentials to attackers.[12] This implies that AI can provide hallucinated content to users and is not yet mature enough to operate autonomously without human intervention. So, human-in-the-loop, human-on-the-loop, and human-over-the-loop are currently necessary. Together, these safeguards are particularly important in defence and national-security environments, where hallucinated or manipulated AI outputs could lead to incorrect intelligence assessments, faulty operational decisions or unintended cyber responses.[13]

The Claude incident[14] reveals a governance problem: a powerful AI system operating in a poorly bounded environment can create harm without malicious intent. For defence organisations, a testing agent that can access the internet, build code, interact with repositories or use operational credentials should be treated as a high-risk insider capability. The distinction between a simulation, a red-team exercise and a real network must be technically enforced rather than assumed from prompts.

The increased dependence on semiconductor suppliers, software vendors, cloud providers, telecom operators, and other third-party entities has increased the possibility of AI agentic-powered attacks on defence supply chains. An attacker could leverage AI capabilities to exploit vulnerabilities and obtain access to operational military data.

Recommendations for Defence Infrastructure

Defence and critical-infrastructure systems should establish AI-powered cyber defence systems that apply the confidentiality, integrity and availability (CIA) triad[15] to AI models, training data, software supply chains and operational networks. Confidentiality keeps information accessible only to authorised people or systems. Integrity ensures data and systems are accurate, unaltered, and detectable if changed. Availability means systems and data are accessible to authorised users when they need them, even during attacks or outages. The implementation of the CIA triad helps to protect both physical and digital ICT infrastructure.

At the same time, it is important to recognise that cyber offence achieves an inherent advantage over cyber defence.[16] An attacker may require only one exploitable vulnerability, misconfiguration or compromised credential to gain initial access to a network, whereas defenders must continuously identify, prioritise and mitigate weaknesses across the entire attack surface.[17] AI

can intensify this asymmetry by accelerating reconnaissance, vulnerability discovery and exploit development.

Defence organisations should establish a continuous supply-chain risk-management programme for semiconductor, software, cloud and telecommunication vendors. All critical suppliers should undergo security vetting, periodic audits, and continuous vulnerability monitoring. Also mandate Software, Hardware, and AI Bills of Materials to map component origin, dependencies and potential exposure.[18] This also led to a requirement of resilience by design with zero trust.

Simulating AI-enabled cyberattacks within a controlled network environment, identifying vulnerabilities and implementing remediation measures is therefore an important means of strengthening under the CIA triad. CERT-In (Computer Emergency Response Team – India) recommends a continuous, intelligence-driven and resilience-oriented approach to defending against AI-assisted cyberattacks, supported by threat monitoring, rapid vulnerability management, security validation and incident-response preparedness.[19] Any critical, high-severity, zero-day, or actively exploited vulnerabilities in defence equipment or software detected by the OEM (Original Equipment Manufacturers) and technology providers should be immediately notified to CERT-In and the designated defence authority.[20]

Defence and national-security organisations should adopt a structured human-oversight framework for all AI-enabled intelligence, cyber defence and decision-support systems.[21] Maintain human approval for high-impact actions such as disabling operational systems, revoking access, or taking defensive action in critical infrastructure.

Notes:

[1] Nilza Amaral, *"The Iran war highlights the creeping use of AI in warfare,"* Chatham House, March 27, 2026, https://www.chathamhouse.org/2026/03/iran-war-highlights-creeping-use-ai-warfare;; Margarita Konaev, "Tomorrow's Technology in Today's War: The Use of AI and Autonomous Technologies in the War in Ukraine and Implications for Strategic Stability," Center for Naval Analysis, September 2023, pp. 1-34, https://www.cna.org/reports/2023/10/Use-of-AI-and-Autonomous-Technologies-in-the-War-in-Ukraine.pdf, Accessed on August 26, 2026.

[2] "What is Malware?," *Microsoft Security,* https://www.microsoft.com/en-in/security/business/security-101/what-is-malware. Accessed on August 21, 2026.

[3] Basabi Pandey, et. al., "AI-Driven Cybercrime and Autonomous Malware: Technological Threats, Legal Challenges and the Need for a New Cybersecurity Framework*," Ijraset Journal for Research in Applied Science and Engineering Technology,* August 17, 2026, https://www.ijraset.com/research-paper/ai-driven-cybercrime-and-autonomous-malware-technological-threats. Accessed on August 25, 2026

[4] D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey," *IEEE Access*, vol. 13, January 29, 2025, pp. 18912-18936, https://ieeexplore.ieee.org/abstract/document/10849561?utm_source=copilot.com. Accessed on August 25, 2026.

[5] Ibid, p.18915.

[6] "Defending Against Frontier AI Driven Cyber Risks," *CERT-In Advisory,* April 26, 2026, https://www.cert-in.org.in/s2cMainServlet?VLCODE=CIAD-2026-0020&pageid=PUBVLNOTES02. Accessed on August 28, 2026.

[7] Catherine A. Theohary and Kelley M. Sayler, "Agentic Artificial Intelligence and Cyberattacks," Congressional Research Service. United States of America, July 06, 2026, https://crsreports.congress.gov. Accessed on August 13, 2026.

[8] "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign," *Anthropic,* November 17, 2025, pp. 2-14, https://www-cdn.anthropic.com/d7dd50dd1185f59be051b307150d877f2b82bd2c.pdf. Accessed on August 13, 2026.

[9] "Disrupting Malicious Uses of AI," *OpenAI* June 2025, pp. 2-46, https://cdn.openai.com/threat-intelligence-reports/5f73af09-a3a3-4a55-992e-069237681620/disrupting-malicious-uses-of-ai-june-2025.pdf#page=4&zoom=100,72,300. Accessed on August 14, 2026.

[10] Ibid, pp. 2-46

[11] "Investigating Three Real-World Incidents in Our Cybersecurity Evaluations," *Anthropic,* July 30, 2026, https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals. Accessed on August 14, 2026.

[12] Ibid.

[13] National Institute of Standards and Technology, United States of America, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," January 2023, pp 4-48, https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf. Accessed on August 26, 2026.

[14] Anthropic, n.11.

[15] National Institute of Standards and Technology, United States of America, "Data Integrity: Detecting and Responding to Ransomware and Other Destructive Events," December 2020, https://www.nccoe.nist.gov/publication/1800-26/VolA/index.html. Accessed on August 26, 2026.

[16] Fernando Montenegro, "Why AI Learned to Attack Before It Learned to Defend," *Futurum*, August 24, 2026, https://futurumgroup.com/insights/why-ai-learned-to-attack-before-it-learned-to-defend/. Accessed on August 26, 2026.

[17] Zach Church, "AI Cyberattacks and Three Pillars for Defense," *MIT Sloan,* September 08, 2025, https://mitsloan.mit.edu/ideas-made-to-matter/ai-cyberattacks-three-pillars-defense. Accessed on August 19, 2026.

[18] Ibid.

[19] Indian Computer Emergency Response Team (CERT-In), Ministry of Electronics and Information Technology, Government of India, "Blueprint for Reducing Exposure and Defending against AI-Assisted Vulnerabilities Exploitation in Digital Infrastructure," May 25, 2026, pp. 1-38, https://www.cert-in.org.in/PDF/Blueprint_for_Defending_against_AI_Assisted_Exploitataion.pdf. Accessed on August 18, 2026.

[20] Indian Computer Emergency Response Team (CERT-In), Ministry of Electronics and Information Technology, Government of India, "Guidelines regarding AI-Accelerated Vulnerability Protection and Response Requirements for Original Equipment Manufacturers (OEMs), and Technology Providers," https://www.cert-in.org.in/s2cMainServlet?pageid=GUIDLNVIEW02&refcode=CISG-2026-03. Accessed on August 26, 2026.

[21] Michael Klare, "AI Plays Major Role in the War on Iran," *Arms Control Association,* May 2026, https://www.armscontrol.org/act/2026-05/news/ai-plays-major-role-war-iran. Accessed on August 26, 2026.