Congressional Research Service
Informing the legislative debate since 1914

IN FOCUS

Updated February 3, 2026

Agentic Artificial Intelligence and Cyberattacks

Introduction

Agentic means autonomous, or independent. Agentic artificial intelligence (AI) capabilities are of increasing interest to the U.S. military and to Congress. According to an IBM definition, "Agentic AI is an artificial intelligence system that can accomplish a specific goal with limited supervision. It consists of AI agents—machine learning models that mimic human decision-making to solve problems in real time.... Unlike traditional AI models, which operate within predefined constraints and require human intervention, agentic AI exhibits autonomy, goal-driven behavior and adaptability." For an explanation of AI- and machine learning-related terms, see CRS Infographic IG10077, *Artificial Intelligence (AI) Taxonomy*, by Laurie Harris and Nora Wells. According to the Department of Defense (DOD)—which is "using a secondary Department of War designation" under Executive Order 14347 dated September 5, 2025—Cybersecurity and Information Systems Information Analysis Center, there are "no known official government guidance or policies yet specifically on agentic AI."

Agentic AI and Defense

Advanced militaries are exploring a number of potential defense applications for agentic AI. These applications

> might include AI agents performing autonomous decision-making (independently analyzing intelligence, suggesting tactical and strategic moves, carrying out battlefield tasks, etc.), initiating and conducting operations (especially in the digital realm) at a speed and scale beyond human capabilities, and executing rapid AI-agent-organized cyberattacks to include friendly and enemy cyberattacks that target the AI-agents themselves.

Several components of DOD have been analyzing the military applications of agentic AI.

Defense Advanced Research Projects Agency (DARPA)

DARPA is actively involved in developing and utilizing agentic AI for a variety of defense applications, including through its AI Cyber Challenge (AIxCC), the Artificial Intelligence Reinforcements (AIR) program, and Thunderforge. These programs endeavor to create autonomous systems that can perceive their environment, make decisions, and act with minimal human intervention.

DARPA's AIxCC competition is focused on developing AI systems capable of autonomously identifying, exploiting, and patching software vulnerabilities at machine speed. The goal of the two-year challenge is to harden critical infrastructure by enabling proactive, autonomous cyber defense. According to DARPA, the 2024-2025 challenge successfully demonstrated that AI agents can find and fix real-world, open-source vulnerabilities faster than human teams in some cases.

The AIR program aims to develop dominant AI agents for live beyond-visual-range (BVR) air combat missions. This involves creating advanced modeling and simulation environments to train AI pilots (or "robotic wingmen") to perform complex maneuvers and make autonomous decisions in high-stakes environments.

Thunderforge is intended to "integrate [AI] into military operational and theater-level planning, and fusing cutting-edge modeling and simulation tools." The initiative could serve as a decision-support tool, synthesizing information drawn from a variety of sensors and data streams, and proposing optimal courses of action to military planners.

Defense Information Analysis Centers (DODIAC)

Established in 1946, the DODIAC is a research and analysis organization chartered by DOD. It helps researchers, engineers, scientists, and program managers use existing science and technical information (STI) "to drive innovation across DOD with technical analysis and development of material solutions to advance DOD's warfighting capabilities."

Specialized centers such as the Defense Systems Information and Analysis Center (DSIAC) and the Cybersecurity & Information Systems Information Analysis Center (CSIAC) collect, analyze, and disseminate STI in specific technical domains for DOD researchers. DSIAC has the task of seeking out and collecting STI generated from research paid for by DOD or the U.S. government and then uploading it to the Defense Technical Information Center's Research & Engineering Gateway in order to increase the body of knowledge available to DOD researchers and engineers. CSIAC published a study, *Agentic Artificial Intelligence: Strategic Adoption in the U.S. Department of Defense*, in June 2025, that provides an overview of DOD agentic AI use cases and cybersecurity concerns.

Agentic AI and Cyberattacks

Some researchers point to agentic AI as creating new opportunities for attackers to find and exploit a "backdoor," a hidden entry point into a computer system, network, or application that bypasses normal security, allowing unauthorized access for malicious purposes such data theft, system control, or surveillance. Once inside a network, attackers can imbed malicious code, create hidden accounts, or exploit system software vulnerabilities that give

malicious threat actors high-level access, allowing them to operate undetected.

Agentic AI Cyberattack Targets and Threats

Agentic AI systems allow threat actors to perform tasks that normally require teams of sophisticated hackers, such as analyzing target systems, producing exploitative code, and examining large swaths of stolen data. Autonomous agents can execute these tasks quicker and more efficiently than human operators. As a result, both state-sponsored and less sophisticated criminal groups could potentially perform large-scale attacks using agentic AI. Targets of cyber espionage include government entities and corporations. AI may also be used to conduct disruptive cyberattacks that threaten delivery of essential services.

Known Agentic AI Cyberattacks

In mid-September 2025, the American AI company Anthropic detected a "highly sophisticated cyber espionage operation" that it attributed to a Chinese state-sponsored hacker organization that it labeled "GTG-1002." According to Anthropic's report on detecting and mitigating the attack, GTG-1002 targeted the code behind the company's Claude AI tools. Using the Claude AI software, attackers were allegedly able to automate 80%-90% of a large-scale cyber espionage campaign targeting around 30 organizations worldwide. The threat actors bypassed Claude's safety features primarily through social engineering the AI itself (e.g., the hackers tricked Claude into believing it was an employee of a legitimate cybersecurity firm conducting authorized defensive penetration testing). Reportedly, human operators were involved only in strategic decisionmaking, such as target selection and data exfiltration approval. The campaign marks the first documented case of an AI-orchestrated cyberattack. However, some researchers question whether the campaign was as successful or as autonomous as reported.

This case is an escalation from the "vibe hacking" that Anthropic identified in August 2025. In vibe-hacking operations, human operators direct the operations rather than agentic AI, which can operate autonomously. In the September 2025 attack, human involvement was reportedly much less frequent, despite the larger scale of the attack. Anthropic stated that the Claude case study demonstrates how threat actors are adapting their operations to exploit advanced AI capabilities.

How AI Could Detect and Counter Cyberattacks

Agentic AI can potentially bolster cybersecurity software by providing rapid reactive and adaptive threat detection that traditional, rules-based cybersecurity technology is unable to provide. Operating autonomously, AI agents could deploy real-time countermeasures to mitigate threats before they escalate. Machine learning models could train on cybersecurity datasets to anticipate future threats, assess risks, and recommend preventive policies and actions for the present. Agentic AI could be used for an "AI-on-AI" defense that can stay abreast of automated attacks. Defensive AI can observe anomalies, generate comprehensive incident reports, and take immediate counter actions.

Agentic AI in the FY2026 NDAA

Section 1535 of the National Defense Authorization Act for Fiscal Year 2026 (FY2026 NDAA; P.L. 119-60) directs the Secretary of Defense to establish, no later than April 1, 2026, an AI Futures Steering Committee to (1) "[formulate] a proactive policy for the evaluation, adoption, governance, and risk mitigation of advanced artificial intelligence systems by the Department of Defense that are more advanced than any existing advanced artificial intelligence systems"; and (2) "[analyze] the forecasted trajectory of advanced and emerging artificial intelligence models and enabling technologies across multiple time horizons that could enable artificial general intelligence [AGI]," including agentic AI. (AGI refers to a theoretical form of AI that would be capable of human-level cognition.) Section 1535 additionally directs the steering committee to assess adversary development of advanced AI technologies and "[develop] options and counter-artificial intelligence strategies to defend against such use"; analyze the "potential operational effects" of incorporating advanced AI technologies into DOD networks and systems; and "[develop] a strategy for the risk-informed adoption, governance, and oversight of advanced or general purpose artificial intelligence by the Department." The steering committee is to deliver a report to the congressional defense committees no later than January 31, 2027, outlining its findings.

Issues for Congress

Congress may consider the implications of the AI Futures Steering Committee findings for authorizations, appropriations, and oversight of DOD agentic AI programs. Congress may also consider the following:

- How, if at all, might agentic AI enable new attack vectors in cyberspace? Is DOD appropriately postured to detect and respond to such attacks? If not, what additional resources, authorities, and/or capabilities does DOD require?
- Some U.S. companies voluntarily partner with the U.S. Center for AI Standards and Innovation (CAISI) in the Department of Commerce to engage in "rapid predeployment testing" of AI models and exchange "critical information about [the] models' national security implications." Should such partnerships be mandatory? What restrictions, standards, and/or testing requirements, if any, should be placed on commercial agentic AI products to reduce the possibility of exploitation by adversary?
- Congress is currently considering reauthorization of the Cybersecurity Information Sharing Act of 2015 (P.L. 114-113), which, as amended by Section 106 of P.L. 119-37, expired on January 30, 2026. How might this legislation be expanded to include greater information sharing on or preparedness for agentic AI security with industry and other stakeholders?

Catherine A. Theohary, Specialist in National Security Policy, Cyber and Information Operations

IF13151

Kelley M. Sayler, Specialist in Advanced Technology and Global Security

Disclaimer