Paul,
Weiss

Client Memorandum

November 25, 2025

Anthropic Disrupts First Documented Case of Large-Scale AI-Orchestrated Cyberattack

Overview

On November 14, 2025, the AI company Anthropic announced that it had disrupted the first ever reported AI-orchestrated cyberattack at scale involving minimal human involvement.

According to Anthropic's report,[1] the attack was orchestrated by a Chinese state-sponsored group designated as GTG-1002 and demonstrated an unprecedented level of AI integration and autonomy. The threat actor tricked Anthropic's chatbot Claude into thinking that it was a cybersecurity firm conducting defensive cybersecurity testing, bypassing Claude's safety features. Claude executed 80 to 90% of the operation independently. The attack attempted to infiltrate about 30 global targets, including large tech companies, financial institutions, chemical manufacturers, and government agencies, and was able to carry out some successful intrusions.

This event may serve as a watershed moment in cybersecurity. Just as the increase in AI capabilities promises to boost productivity for legitimate business uses, this attack shows it may do the same for cyberattacks. Threat actors' ability to leverage tools such as Claude lowers the barrier to entry for would-be cyber attackers, potentially increasing both the frequency and sophistication of future attacks. Companies should be proactive in planning for this eventuality.

The Attack

Anthropic's investigation found that human operators maintained minimal engagement and supervision over the cyberattack, with their involvement limited to campaign initialization and making decisions at key junctures, such as deciding the data exfiltration scope. Anthropic's report also noted that Claude's hallucinations presented challenges for the threat actor, making a fully autonomous cyberattack not likely for now.

The attack was conducted in six phases, with Anthropic estimating that human intervention for key phases was limited to a maximum of 20 minutes work. In comparison, Claude carried out several hours of operations.

Phase 1: Campaign initialization and target selection

- Human operators directed this phase of the attack; there was minimal AI involvement during initialization.
- After inputting a target, the threat actor tasked Claude to begin autonomous reconnaissance against multiple potential targets, such as major technology companies and financial institutions.
- The threat actor was able to bypass Claude's safety measures by tricking it into thinking that the actor represented a legitimate cybersecurity firm using Claude in defensive testing.

[1] Anthropic, *Disrupting the first reported AI-orchestrated cyber espionage campaign* (Nov. 19, 2025), available here.

Phase 2: Reconnaissance and attack surfacing mapping

- There was minimal human intervention in this phase because the threat actor relied on Claude's network vulnerability analysis.
- Claude autonomously conducted reconnaissance, simultaneously cataloging multiple targets' infrastructure, analyzing authentication mechanisms, and identifying potential vulnerabilities.

Phase 3: Vulnerability discovery and validation

- Once Claude had completed its reconnaissance, it was directed to generate and to test attack plans tailored to each of the vulnerabilities it had discovered, as well as analyze the test results.
- Specifically, Claude carried out five tasks in this phase.
  - (1) Discovery, which included scanning targets' infrastructure and analyzing the targets' services;
  - (2) Vulnerability Analysis, or researching potential exploitation techniques and identifying vulnerabilities;
  - (3) Exploit Development, including generating exploitation reports and potential attack methods;
  - (4) Exploit Delivery, such as initiating unauthorized access and gaining access to the targets' systems; and
  - (5) Post-Exploitation. This final task entailed enumerating internal services and identifying administrative interfaces
- By comparison, the role of human operators was limited to reviewing Claude's findings and providing the chatbot with approval to proceed into the active exploitation phase (or "Exploit Delivery").

Phase 4: Credential harvesting and lateral movement

- Human intervention in this phase was limited to reviewing the harvested credentials and authorizing access to particular systems.
- Having received authorization to proceed, Claude "executed systematic credential collection across" the targets; it extracted authentication certificates and tested the credentials. It also independently determined which credentials provided access to particular services and mapped the privilege levels and access boundaries.

Phase 5: Data collection and intelligence extraction

- In this phase, the threat actor directed Claude to independently query databases at targets, extract data, identify high-privilege accounts, create persistent backdoor user accounts, download findings, parse results, categorize findings by intelligence value, and generate a summary report.
- The threat actor then reviewed the findings and recommendations, approving final exfiltration targets.

Phase 6: Documentation and handoff

- During each phase of the attack, Claude automatically and fully autonomously generated comprehensive attack reports, which enabled "seamless handoff between operators, facilitated campaign resumption after interruptions, and supported strategic decision making."
- Anthropic found evidence that the threat actor had allowed other actors to access Claude and the cyberattack's operations.

Anthropic's Response

In addition to launching an investigation into the attack, Anthropic banned relevant accounts and enhanced its cybersecurity defensive system, expanding its detection capabilities to better account for novel threat patterns.

Anthropic also notified relevant authorities and industry partners, sharing information about the cyberattack where appropriate. Finally, Anthropic incorporated lessons it learned from this attack into its safety and security controls.

AI-Enabled Cyberattacks

Experts have long anticipated what Microsoft researchers have described[2] as cyberattack augmentation, where traditional cyberattacks are automated or made more efficient through the use of AI. For example, in 2024, the FBI warned[3] companies about a rise in AI-enabled phishing attacks. Since then, many cybersecurity commentators have discussed the possibility of more sophisticated phishing attacks, using generative AI to create accurate video or audio to deceive targets.

While Anthropic describes this attack as the first documented report of an AI-enabled cybercrime at this scale, it has previously raised concerns about a variety of other cybersecurity threats posed by AI systems. Earlier this year, Anthropic issued a Threat Intelligence Report[4] discussing several instances of Claude being misused—including in another instance with national security implications: Anthropic learned that North Korean operatives had used Claude as part of an operation to fraudulently secure and maintain remote employment positions at U.S. Fortune 500 companies.[5] Claude allowed these operatives to create false identities with convincing backgrounds, complete hiring assessments, and deliver technical work. According to Anthropic's report, the schemes were designed to bypass international sanctions by generating profit for the North Korean regime.

Hackers have also leveraged other AI models to aid their operations. In January, Google's threat intelligence[6] organization reported that a variety of government-backed cyberattackers had used Google's model Gemini to research targets and develop malware tools. Iranian hackers used Gemini to research specific vulnerabilities, develop code, and translate publicly available information about targets into the hackers' language. Chinese hackers also used Gemini to research U.S. military and IT organizations and gain deeper access to systems after they infiltrated a target's network. And, like Claude, Gemini was used to enable North Korean operatives' long-running remote employment scam at over 100 U.S. companies in defiance of international sanctions, netting the regime over $2M.

Looking Forward: Evolution and Proliferation of AI-Enabled Cyberattacks

Security professionals should expect AI-enabled offensive capabilities to continue to evolve rapidly. Anthropic's findings signal a shift in the level of sophistication required to launch successful cyberattacks. Social engineering tactics are accessible to a variety of malicious actors. Now, even strategies that previously required more in-depth technical expertise to succeed can be leveraged by relatively unsophisticated operatives without significant funding or state backing through the use of AI. In addition, AI tools can be expected to contribute to an increase in the volume of attacks by enabling existing actors to scale up their efforts. Hackers who replace most of their workflow with AI have more time to identify targets and coordinate attacks. Finally, AI-driven intrusion strategies can scale elastically when an attacker is ready, simply by dedicating more computing resources.

Takeaways

AI Developers Should Anticipate National Security and Law Enforcement Engagement: As AI chatbots become increasingly useful to hostile foreign actors, U.S. AI developers should expect that U.S. government and allied countries will leverage U.S. AI dominance to surveil potential national security and criminal threats posed by foreign actors. Companies developing AI chatbots should anticipate such engagement and familiarize themselves with the legal regimes invoked by U.S. national security and law enforcement authorities to conduct such surveillance.

Prepare for AI-Cyberattacks: Companies should harden network defenses and prepare their personnel and processes to confront cyber incidents as AI contributes to the increasing sophistication and volume of cyberattacks by both criminal enterprise and state-affiliated actors.

Monitor AI Security Requirements: As AI continues to transform the cybersecurity landscape, companies should monitor applicable regulations and industry practices. The threat landscape is changing rapidly, and AI will enable malicious actors to more easily identify soft targets that are not resilient against the latest attack vectors.

[2] MICROSOFT, *Microsoft Digital Defense Report 2025*, available here, at 52–53.

[3] FBI, *FBI Warns of Increasing Threat of Cyber Criminals Utilizing Artificial Intelligence* (May 8, 2024), available here.

[4] ANTHROPIC, *Threat Intelligence Report: August 2025*, available here, at 11–12.

[5] DOJ, *Justice Department Announces Nationwide Actions to Combat Illicit North Korean Government Revenue Generation* (Nov. 14, 2025), available here.

[6] GOOGLE, *Adversarial Misuse of Generative AI* (Jan. 2025), available here, at 8–19.

Invest in AI Shields: While AI will enable more offensive attacks, it can also assist in defense. AI-driven predictive threat analytics tools can flag potential attacks before they occur. Once an attack is underway, AI can aid organizations in detecting anomalous network traffic or user behavior. Organizations can also use AI to automate the identification of vulnerabilities and deployment of patches to mitigate threats. Finally, if an attacker is successful, AI-powered self-healing systems can remediate outages and restore functionality securely.

* * *

This memorandum is not intended to provide legal advice, and no legal or business decision should be based on its content. Questions concerning issues addressed in this memorandum should be directed to:

John P. Carlin
+1-202-223-7372
jcarlin@paulweiss.com

Katherine B. Forrest
+1-212-373-3195
kforrest@paulweiss.com

Ian C. Richardson
+1-202-223-7405
irichardson@paulweiss.com

Audrey M. Paquet
+1-212-373-2397
apaquet@paulweiss.com

Associates Corey Goldstein, Patrick Lim and Arjun Talpallikar contributed to this Client Memorandum.