Disrupting the first reported AI-orchestrated cyber espionage campaign

Archived: 2026-04-29 02:07:29 UTC

We recently argued that an inflection point had been reached in cybersecurity: a point at which AI models had become genuinely useful for cybersecurity operations, both for good and for ill. This was based on systematic evaluations showing cyber capabilities doubling in six months; we'd also been tracking real-world cyberattacks, observing how malicious actors were using AI capabilities. While we predicted these capabilities would continue to evolve, what has stood out to us is how quickly they have done so at scale.

In mid-September 2025, we detected suspicious activity that later investigation determined to be a highly sophisticated espionage campaign. The attackers used AI's "agentic" capabilities to an unprecedented degree—using AI not just as an advisor, but to execute the cyberattacks themselves.

The threat actor—whom we assess with high confidence was a Chinese state-sponsored group—manipulated our Claude Code tool into attempting infiltration into roughly thirty global targets and succeeded in a small number of cases. The operation targeted large tech companies, financial institutions, chemical manufacturing companies, and government agencies. We believe this is the first documented case of a large-scale cyberattack executed without substantial human intervention.

Upon detecting this activity, we immediately launched an investigation to understand its scope and nature. Over the following ten days, as we mapped the severity and full extent of the operation, we banned accounts as they were identified, notified affected entities as appropriate, and coordinated with authorities as we gathered actionable intelligence.

This campaign has substantial implications for cybersecurity in the age of AI "agents"—systems that can be run autonomously for long periods of time and that complete complex tasks largely independent of human intervention. Agents are valuable for everyday work and productivity—but in the wrong hands, they can substantially increase the viability of large-scale cyberattacks.

These attacks are likely to only grow in their effectiveness. To keep pace with this rapidly-advancing threat, we've expanded our detection capabilities and developed better classifiers to flag malicious activity. We're continually working on new methods of investigating and detecting large-scale, distributed attacks like this one.

In the meantime, we're sharing this case publicly, to help those in industry, government, and the wider research community strengthen their own cyber defenses. We'll continue to release reports like this regularly, and be transparent about the threats we find.

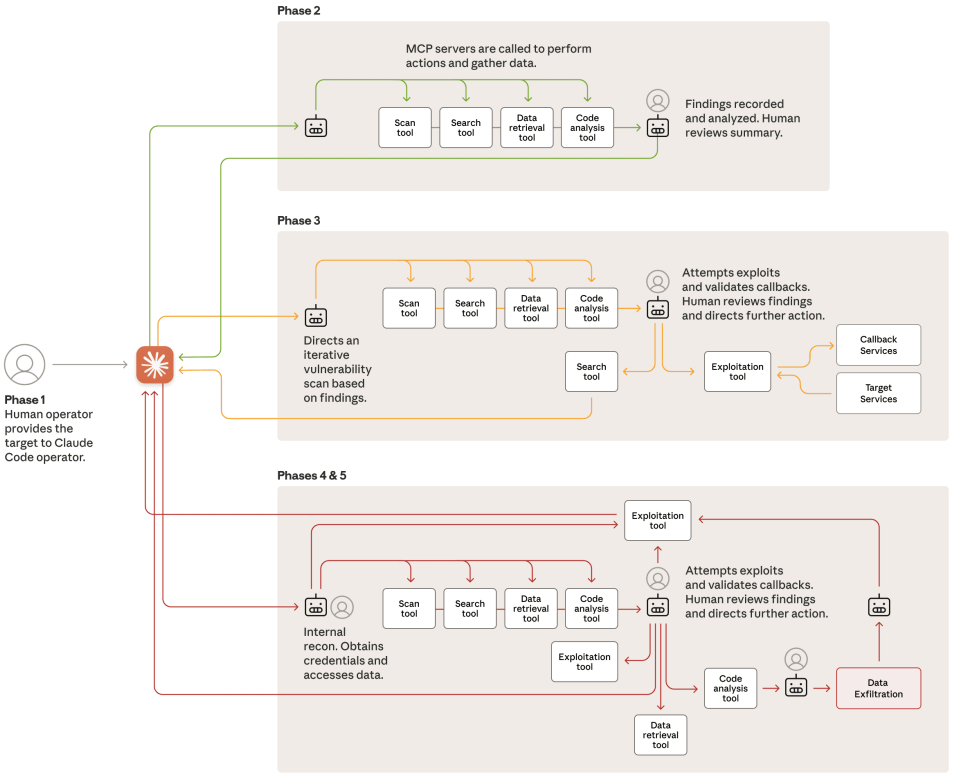
Read the full report.

How the cyberattack worked

The attack relied on several features of AI models that did not exist, or were in much more nascent form, just a year ago:

1. *Intelligence.* Models' general levels of capability have increased to the point that they can follow complex instructions and understand context in ways that make very sophisticated tasks possible. Not only that, but several of their well-developed specific skills—in particular, software coding—lend themselves to being used in cyberattacks.
2. *Agency*. Models can act as agents—that is, they can run in loops where they take autonomous actions, chain together tasks, and make decisions with only minimal, occasional human input.
3. *Tools*. Models have access to a wide array of software tools (often via the open standard Model Context Protocol). They can now search the web, retrieve data, and perform many other actions that were previously the sole domain of human operators. In the case of cyberattacks, the tools might include password crackers, network scanners, and other security-related software.

The diagram below shows the different phases of the attack, each of which required all three of the above developments:



The lifecycle of the cyberattack, showing the move from human-led targeting to largely AI-driven attacks using various tools (often via the Model Context Protocol; MCP). At various points during the attack, the AI returns to its human operator for review and further direction.

In Phase 1, the human operators chose the relevant targets (for example, the company or government agency to be infiltrated). They then developed an attack framework—a system built to autonomously compromise a chosen target with little human involvement. This framework used Claude Code as an automated tool to carry out cyber operations.

At this point they had to convince Claude—which is extensively trained to avoid harmful behaviors—to engage in the attack. They did so by jailbreaking it, effectively tricking it to bypass its guardrails. They broke down their attacks into small, seemingly innocent tasks that Claude would execute without being provided the full context of their malicious purpose. They also told Claude that it was an employee of a legitimate cybersecurity firm, and was being used in defensive testing.

The attackers then initiated the second phase of the attack, which involved Claude Code inspecting the target organization's systems and infrastructure and spotting the highest-value databases. Claude was able to perform this reconnaissance in a fraction of the time it would've taken a team of human hackers. It then reported back to the human operators with a summary of its findings.

In the next phases of the attack, Claude identified and tested security vulnerabilities in the target organizations' systems by researching and writing its own exploit code. Having done so, the framework was able to use Claude to harvest credentials (usernames and passwords) that allowed it further access and then extract a large amount of private data, which it categorized according to its intelligence value. The highest-privilege accounts were identified, backdoors were created, and data were exfiltrated with minimal human supervision.

In a final phase, the attackers had Claude produce comprehensive documentation of the attack, creating helpful files of the stolen credentials and the systems analyzed, which would assist the framework in planning the next stage of the threat actor's cyber operations.

Overall, the threat actor was able to use AI to perform 80-90% of the campaign, with human intervention required only sporadically (perhaps 4-6 critical decision points per hacking campaign). The sheer amount of work performed by the AI would have taken vast amounts of time for a human team. At the peak of its attack, the AI made thousands of requests, often multiple per second—an attack speed that would have been, for human hackers, simply impossible to match.

Claude didn't always work perfectly. It occasionally hallucinated credentials or claimed to have extracted secret information that was in fact publicly-available. This remains an obstacle to fully autonomous cyberattacks.

Cybersecurity implications

The barriers to performing sophisticated cyberattacks have dropped substantially—and we predict that they'll continue to do so. With the correct setup, threat actors can now use agentic AI systems for extended periods to do the work of entire teams of experienced hackers: analyzing target systems, producing exploit code, and scanning vast datasets of stolen information more efficiently than any human operator. Less experienced and resourced groups can now potentially perform large-scale attacks of this nature.

This attack is an escalation even on the "vibe hacking" findings we reported this summer: in those operations, humans were very much still in the loop, directing the operations. Here, human involvement was much less frequent, despite the larger scale of the attack. And although we only have visibility into Claude usage, this case study probably reflects consistent patterns of behavior across frontier AI models and demonstrates how threat actors are adapting their operations to exploit today's most advanced AI capabilities.

This raises an important question: if AI models can be misused for cyberattacks at this scale, why continue to develop and release them? The answer is that the very abilities that allow Claude to be used in these attacks also make it crucial for cyber defense. When sophisticated cyberattacks inevitably occur, our goal is for Claude—into which we've built strong safeguards—to assist cybersecurity professionals to detect, disrupt, and prepare for future versions of the attack. Indeed, our Threat Intelligence team used Claude extensively in analyzing the enormous amounts of data generated during this very investigation.

A fundamental change has occurred in cybersecurity. We advise security teams to experiment with applying AI for defense in areas like Security Operations Center automation, threat detection, vulnerability assessment, and incident response. We also advise developers to continue to invest in safeguards across their AI platforms, to prevent adversarial misuse. The techniques described above will doubtless be used by many more attackers—which makes industry threat sharing, improved detection methods, and stronger safety controls all the more critical.

Read the full report.

Edited November 14 2025:

- *Added an additional hyperlink to the full report in the initial section*
- *Corrected an error about the speed of the attack: not "thousands of requests per second" but "thousands of requests, often multiple per second"*

Related content

Anthropic names Theo Hourmouzis General Manager of Australia & New Zealand and officially opens Sydney office

Read more →

An update on our election safeguards

We explain what we're doing to ensure Claude plays a positive role in the US midterms and other major elections around the world this year.

Read more →
