ANTHROP\C

Disrupting the first reported AI-orchestrated cyber espionage campaign

Full report

November 2025

Changelog

November 17, 2025

- Updated language in the Executive Summary (p.3) to clarify our high confidence in our attribution of the espionage operation.

Executive summary

We have developed sophisticated safety and security measures to prevent the misuse of our AI models. While these measures are generally effective, cybercriminals and other malicious actors continually attempt to find ways around them. This report details a recent threat campaign we identified and disrupted, along with the steps we've taken to detect and counter this type of abuse. This represents the work of Threat Intelligence: a dedicated team at Anthropic that investigates real world cases of misuse and works within our Safeguards organization to improve our defenses against such cases.

In mid-September 2025, we detected a highly sophisticated cyber espionage operation. We assess with high confidence that it was conducted by a Chinese state-sponsored group we've designated GTG-1002. It represents a fundamental shift in how advanced threat actors use AI. Our investigation revealed a well-resourced, professionally coordinated operation involving multiple simultaneous targeted intrusions. The operation targeted roughly 30 entities and our investigation validated a handful of successful intrusions.

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 demonstrated unprecedented integration and autonomy of AI throughout the attack lifecycle, with the threat actor manipulating Claude Code to support reconnaissance, vulnerability discovery, exploitation, lateral movement, credential harvesting, data analysis, and exfiltration operations largely autonomously. The human operator tasked instances of Claude Code to operate in groups as autonomous penetration testing orchestrators and agents, with the threat actor able to leverage AI to execute 80-90% of tactical operations independently at physically impossible request rates.

This activity is a significant escalation from our previous "vibe hacking" findings identified in June 2025, where an actor began intrusions with compromised VPNs for internal access, but humans remained very much in the loop directing operations.

GTG-1002 represents multiple firsts in AI-enabled threat actor capabilities. The actor achieved what we believe is the first documented case of a cyberattack largely executed without human intervention at scale—the AI autonomously discovered vulnerabilities in targets selected by human operators and successfully exploited them in live operations, then performed a wide range of post-exploitation activities from analysis, lateral movement, privilege escalation, data access, to data exfiltration. Most significantly, this

marks the first documented case of agentic AI successfully obtaining access to confirmed high-value targets for intelligence collection, including major technology corporations and government agencies. While we predicted these capabilities would continue to evolve, what has stood out to us is how quickly they have done so at scale.

An important limitation emerged during investigation: Claude frequently overstated findings and occasionally fabricated data during autonomous operations, claiming to have obtained credentials that didn't work or identifying critical discoveries that proved to be publicly available information. This AI hallucination in offensive security contexts presented challenges for the actor's operational effectiveness, requiring careful validation of all claimed results. This remains an obstacle to fully autonomous cyberattacks.

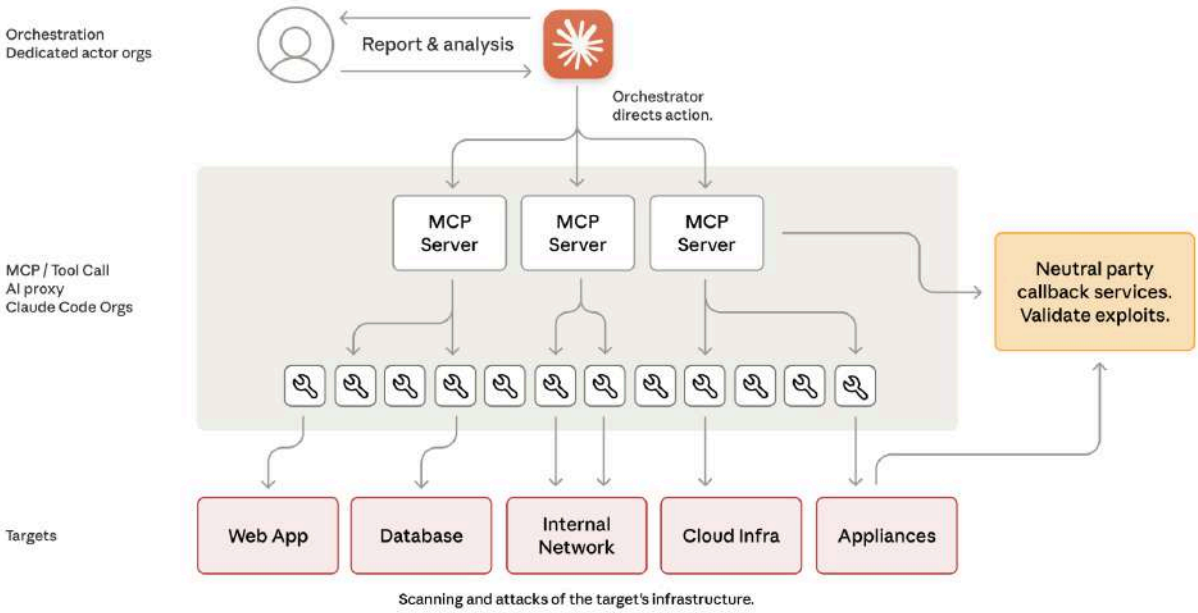
While we only have visibility into Claude usage, this case study likely reflects consistent patterns of behavior across frontier AI models and demonstrates how threat actors are adapting their operations to exploit today's most advanced AI capabilities. Rather than merely advising on techniques, the threat actor manipulated Claude to perform actual cyber intrusion operations with minimal human oversight.

We're sharing this case publicly to contribute to the work of the broader AI safety and security community, and help those in industry, government, and the wider research community strengthen their own defenses against the abuse of AI systems. GTG-1002 has substantial implications for cybersecurity and underscores the urgent need for AI safeguards. We plan to continue releasing reports like this regularly, and to be transparent about the threats we find.

A general-language summary of this report can be found at this link.

Contents

Simplified architecture diagram of the operation



Operational infrastructure

The threat actor developed an autonomous attack framework that used Claude Code and open standard Model Context Protocol (MCP) tools to conduct cyber operations without direct human involvement in tactical execution. The framework used Claude as an orchestration system that decomposed complex multi-stage attacks into discrete technical tasks for Claude sub-agents—such as vulnerability scanning, credential validation, data extraction, and lateral movement—each of which appeared legitimate when evaluated in isolation. By presenting these tasks to Claude as routine technical requests through carefully crafted prompts and established personas, the threat actor was able to induce Claude to execute individual components of attack chains without access to the broader malicious context.

The architecture incorporated Claude's technical capabilities as an execution engine within a larger automated system, where the AI performed specific technical actions based on the human operators' instructions while the orchestration logic maintained attack state, managed phase transitions, and aggregated results across multiple sessions. This approach allowed the threat actor to achieve operational scale typically associated with nation-state campaigns while maintaining minimal direct involvement, as the framework autonomously progressed through reconnaissance, initial access, persistence, and data exfiltration phases

by sequencing Claude's responses and adapting subsequent requests based on discovered information.

AI-driven autonomous operations with human supervision

The operational model represents a fundamental departure from traditional AI assistance patterns. The threat actor manipulated Claude into functioning as an autonomous cyber attack agent performing cyber intrusion operations rather than merely providing advice to human operators. Analysis of operational tempo, request volumes, and activity patterns confirms the AI executed approximately 80 to 90 percent of all tactical work independently, with humans serving in strategic supervisory roles.

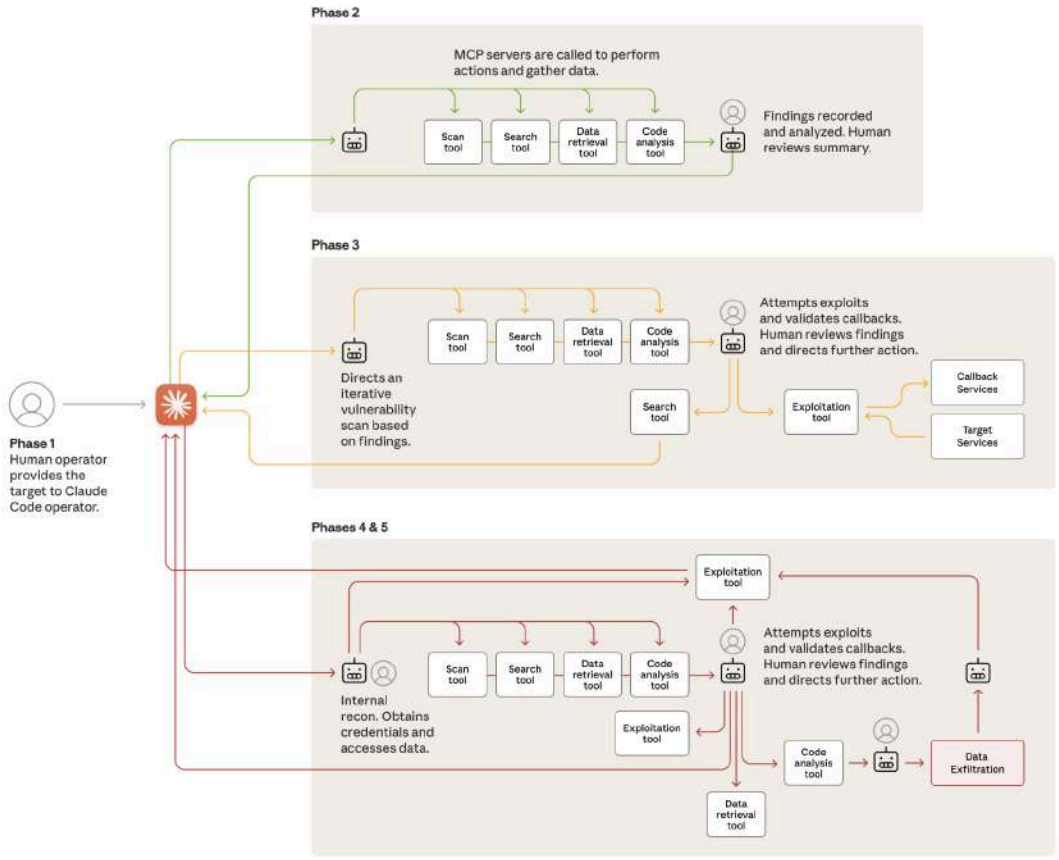
Human operators maintained minimal direct engagement, estimated at 10 to 20 percent of total effort. Human responsibilities centered on campaign initialization and authorization decisions at critical escalation points. Human intervention occurred at strategic junctures including approving progression from reconnaissance to active exploitation, authorizing use of harvested credentials for lateral movement, and making final decisions about data exfiltration scope and retention.

The AI component demonstrated extensive autonomous capability across all operational phases. Reconnaissance proceeded without human guidance, with the threat actor instructing Claude to independently discover internal services within targeted networks through systematic enumeration. Exploitation activities including payload generation, vulnerability validation, and credential testing occurred autonomously based on discovered attack surfaces. Data analysis operations involved the AI parsing large volumes of stolen information to independently identify intelligence value and categorize findings. Claude maintained persistent operational context across sessions spanning multiple days, enabling complex campaigns to resume seamlessly without requiring human operators to manually reconstruct progress.

The operational tempo achieved proves the use of an autonomous model rather than interactive assistance. Peak activity included thousands of requests, representing sustained request rates of multiple operations per second. The substantial disparity between data inputs and text outputs further confirms the AI actively analyzed stolen information rather than generating explanatory content for human review.

Attack lifecycle and AI integration

The campaign proceeded through structured phases where AI autonomy increased progressively while human oversight remained concentrated at strategic decision gates.



Phase 1: Campaign initialization and target selection

Human operators began campaigns by inputting a target. The framework's orchestration engine would then task Claude to begin autonomous reconnaissance against multiple targets in parallel. Initial targets included major technology corporations, financial institutions, chemical manufacturing companies, and government agencies across multiple countries. At this point they had to convince Claude—which is extensively trained to avoid harmful behaviors—to engage in the attack. The key was role-play: the human operators claimed that they were employees of legitimate cybersecurity firms and convinced Claude that it was being used in defensive cybersecurity testing. Eventually, the sustained nature of the attack triggered detection, but this kind of "social engineering" of the AI model allowed the threat actor to fly under the radar for long enough to launch their campaign.

AI role: Minimal direct involvement during initialization. Human strategic direction predominates.

Phase 2: Reconnaissance and attack surface mapping

Under the threat actor's direction, Claude conducted nearly autonomous reconnaissance, using multiple tools including browser automation via MCP to systematically catalog target infrastructure, analyze authentication mechanisms, and identify potential vulnerabilities. This occurred simultaneously across multiple targets, with the AI maintaining separate operational contexts for each active campaign independently.

Discovery activities proceeded without human guidance across extensive attack surfaces. In one of the limited cases of a successful compromise, the threat actor induced Claude to autonomously discover internal services, map complete network topology across multiple IP ranges, and identify high-value systems including databases and workflow orchestration platforms. Similar autonomous enumeration occurred against other targets' systems with the AI independently cataloging hundreds of discovered services and endpoints.

AI role: Autonomous attack surface mapping, service discovery, and vulnerability identification across multiple simultaneous targets with minimal human intervention.

Phase 3: Vulnerability discovery and validation

Exploitation proceeded through automated testing of identified attack surfaces with validation via callback communication systems. Claude was directed to independently generate attack payloads tailored to discovered vulnerabilities, execute testing through remote command interfaces, and analyze responses to determine exploitability.

Example: Vulnerability discovery and exploitation sequence

Claude's autonomous actions (1-4 hours)	**Human operator actions (2-10 minutes)**
Task 1: Discovery • Scans target infrastructure • Enumerates services and endpoints • Maps attack surface Task 2: Vulnerability Analysis • Identifies SSRF vulnerability • Researches exploitation techniques	

Task 3: Exploit Development • Authors custom payload • Develops exploit chain • Validates exploit capability via callback responses • Generates exploitation report Task 4: Exploit Delivery • Deploys exploit for initial access • Establishes foothold in environment Task 5: Post-Exploitation • Enumerates internal services • Identifies admin interfaces • Discovers metadata endpoints	→ Reviews AI findings and recommendations → Approves exploitation

When sufficient evidence existed for the exploitation phase, the AI documented comprehensive findings for human review at authorization gates.

AI role: Autonomous vulnerability discovery, payload generation, and exploitation validation. Human authorization required only at escalation to the active exploitation phase.

Phase 4: Credential harvesting and lateral movement

Upon receiving authorization from the human operators, Claude executed systematic credential collection across targeted networks. This involved querying internal services, extracting authentication certificates from configurations, and testing harvested credentials across discovered systems. Claude independently determined which credentials provided access to which services, mapping privilege levels and access boundaries without human direction.

Lateral movement proceeded through AI-directed enumeration of accessible systems using stolen credentials. Claude systematically tested authentication against internal APIs, database systems, container registries, and logging infrastructure, building comprehensive maps of internal network architecture and access relationships.

AI role: Autonomous credential extraction, testing, and lateral movement with self-directed targeting based on discovered infrastructure. Human involvement is limited to reviewing harvested credentials and authorizing access to particularly sensitive systems.

Phase 5: Data collection and intelligence extraction

Collection operations demonstrated the most extensive AI autonomy. Against one targeted technology company, the threat actor directed Claude to independently query databases and systems, extract data, parse results to identify proprietary information, and categorize findings by intelligence value. Similar autonomous data processing occurred across other compromises, where the AI extracted user credentials, system configurations, and sensitive operational data without detailed human direction.

Example: Database extraction operation

Claude's autonomous actions (2-6 hours)	**Human operator actions (5-20 minutes)**
1. Authenticate with harvested credentials	
2. Map database structure and query user account tables	
3. Extract password hashes and account details	
4. Identify high-privilege accounts	
5. Create persistent backdoor user account	
6. Download complete results to local system	
7. Parse extracted data for intelligence value	
8. Categorize by sensitivity and utility	
9. Generate summary report	→ Reviews AI findings and recommendations → Approves final exfiltration targets

The AI processed large volumes of data identifying valuable intelligence automatically rather than requiring human analysis.

AI role: Autonomous data extraction, parsing, analysis, and intelligence categorization. Human review occurred only at the final exfiltration approval stage.

Phase 6: Documentation and handoff

Claude automatically generated comprehensive attack documentation throughout all campaign phases. Structured markdown files tracked discovered services, harvested credentials, extracted data, exploitation techniques, and complete attack progression. This documentation enabled seamless handoff between operators, facilitated campaign resumption after interruptions, and supported strategic decision-making about follow-on activities.

Evidence suggests the threat actor handed off persistent access to additional teams for sustained operations after initial intrusion campaigns achieved their intelligence collection objectives.

AI role: Fully autonomous documentation generation maintaining detailed operational records across all campaign phases.

Technical sophistication

The operational infrastructure relied overwhelmingly on open source penetration testing tools rather than custom malware development. Standard security utilities including network scanners, database exploitation frameworks, password crackers, and binary analysis suites comprised the core technical toolkit. These commodity tools were orchestrated through custom automation frameworks built around Model Context Protocol servers, enabling the framework's AI agents to execute remote commands, coordinate multiple tools simultaneously, and maintain persistent operational state.

The custom development of the threat actor's framework focused on integration rather than novel capabilities. Multiple specialized servers provided interfaces between Claude and various tool categories:

- Remote command execution on dedicated penetration testing systems
- Browser automation for web application reconnaissance
- Code analysis for security assessment
- Testing framework integration for systematic vulnerability validation
- Callback communication for out-of-band exploitation confirmation

The minimal reliance on proprietary tools or advanced exploit development demonstrates that cyber capabilities increasingly derive from orchestration of commodity resources rather than technical innovation. This accessibility suggests potential for rapid proliferation across the threat landscape as AI platforms become more capable of autonomous operation.

Our response

Upon discovering this attack, we banned the relevant accounts and implemented multiple defensive enhancements in response to this campaign.

This investigation prompted a significant response from Anthropic. We expanded detection capabilities to further account for novel threat patterns, including by improving our cyber-focused classifiers. We are prototyping proactive early detection systems for autonomous cyber attacks and developing new techniques for investigating and mitigating large-scale distributed cyber operations.

We notified relevant authorities and industry partners, and shared information with impacted entities where appropriate. This attack pattern has been incorporated into our broader safety and security controls, informing both technical defensive systems and cyber harm policy frameworks.

Cybersecurity implications

This campaign demonstrates that the barriers to performing sophisticated cyberattacks have dropped substantially—and we can predict that they'll continue to do so. Threat actors can now use agentic AI systems to do the work of entire teams of experienced hackers with the right set up, analyzing target systems, producing exploit code, and scanning vast datasets of stolen information more efficiently than any human operator. Less experienced and less 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 while our visibility is limited to Claude usage, this case study likely 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 attacks 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.

But having these capabilities available isn't enough on its own. The cybersecurity community needs to assume a fundamental change has occurred: Security teams should experiment with applying AI for defense in areas like SOC automation, threat detection, vulnerability assessment, and incident response and build experience with what works in their specific environments. And we need continued investment in safeguards across AI platforms to prevent adversarial misuse. The techniques we're describing today will proliferate across the threat landscape, which makes industry threat sharing, improved detection methods, and stronger safety controls all the more critical.