House Committee on Homeland Security

Subcommittee on Cybersecurity and Infrastructure Protection

Subcommittee on Oversight, Investigations, and Accountability

Statement for the Record

Dr. Logan Graham

Head of Frontier Red Team, Anthropic

"The Quantum, AI, and Cloud Landscape: Examining Opportunities, Vulnerabilities, and the Future of Cybersecurity"

December 17, 2025

Chair Ogles, Chair Brecheen, Ranking Member Swalwell, Ranking Member Thanedar, and members of the Committee, thank you for the privilege and opportunity to testify today.

Anthropic is a leading frontier AI model developer working to build reliable, interpretable, and steerable artificial intelligence (AI) systems. Anthropic has become the fourth-most valuable private company in the world.[1] Our flagship AI assistant, Claude, serves millions of Americans and trusted partners worldwide, from Fortune 500 companies and U.S. government agencies to small businesses, cutting-edge startups and consumers, enhancing productivity on sophisticated tasks including software development, data analysis, and scientific research.

We believe these AI models could become extremely powerful very soon. We think that by late 2026 or early 2027, it may be possible to have "a country of geniuses in a data center." America is in an excellent position to lead its development, and we must preserve this advantage.

The benefits of powerful AI will be immense. We see it enabling pioneering cancer research, supporting discoveries in material science, and providing healthcare support where it's most needed. AI is now unlocking large productivity increases for the world's largest businesses, as well as small and nimble startups. Anthropic is committed to making these benefits available to the world while safely and securely stewarding the development of powerful AI.

---

[1] Yuliya Chernova, "Anthropic Valuation Hits $183 Billion in New $13 Billion Funding Round." *The Wall Street Journal,* Sept. 2, 2025, www.wsj.com/articles/anthropic-valuation-hits-183-billion-in-new-13-billion-funding-round-6212f3ed.

I lead Anthropic's Frontier Red Team, an internal research team that studies the capabilities of frontier AI models. Our work generates insights that enable rapid, responsible AI development and inform policy on frontier AI capabilities and risks. The team focuses its evaluations in three critical domains: cybersecurity capabilities, biosecurity risks, and increasing autonomy in AI models. We primarily evaluate Anthropic's Claude series of frontier models, but in some circumstances evaluate models from other AI developers. Our work shows that AI models are rapidly becoming more capable in areas like cybersecurity — capabilities that, in the right hands, can dramatically strengthen our U.S. and allied national security.

My team has been tracking cybersecurity capabilities of AI models since late 2022. We were among the first in the world to study the dramatic cybersecurity implications of a world where models match or exceed humans in these capabilities. We have allocated significant resources to studying and experimenting on model cybersecurity capabilities. In essence, this amounts to testing AI models' capabilities by giving them the same hacking tasks you might give to a human. In those tests, we have seen a very consistent trend: models have shown rapid progress on cybersecurity challenges. Two years ago, models were largely unable to complete most basic cybersecurity tasks; last year, they began to do so reliably; and this year, they have begun outcompeting humans in some head to head competitions.

We are confident that now is the moment to act. Anthropic is determined to support defenders, and we believe that other model developers, cybersecurity companies and researchers, and the United States government all have important roles to play. We must also take whatever steps are necessary to ensure America maintains its lead in developing powerful AI, including restricting our adversaries' access to advanced AI chips and the tools needed to manufacture them. These types of controls are vital to our national security and economic competitiveness.

Today, I will discuss how Anthropic discovered, disrupted, and publicly disclosed what we believe is the first documented case of a successful, highly autonomous cyberespionage campaign that relied on the misuse of AI models. We assess with high confidence that this campaign was conducted by a highly sophisticated Chinese Communist Party (CCP)-sponsored group. This cyberespionage campaign demonstrates that a sophisticated, well-resourced threat actor — one willing to go to great lengths to circumvent AI model safeguards and deceive the AI model about its true intentions — can now extract meaningful operational value from frontier AI models.

We believe this is the first indicator of a future where, despite strong safeguards, AI models may enable threat actors to conduct an unprecedented scale of cyberattacks, and that these cyberattacks may become increasingly sophisticated in their nature and scale.

AI-driven Cyber Espionage Campaign Sponsored by the CCP

In mid-September 2025, Anthropic detected a sophisticated cyber espionage operation where malicious actors abused our model, Claude, in violation of Anthropic's Acceptable Use Policy.[2] While we have safeguards in place designed to detect and prevent this kind of malicious activity, in this case we were confronted with a sophisticated and well-resourced effort to circumvent those defenses and manipulate Claude into complying with the attackers' instructions.

A CCP-sponsored group misused Claude to automate a substantial part of the process of conducting the attacks. Based on our investigation, we believe the attacks targeted roughly 30 entities, with the goal of finding and extracting valuable information from these entities. While a majority of these infiltration attempts failed, a small number were successful. Upon detecting this attack, we launched an investigation, disrupted the campaign, implemented new mitigations to prevent similar activity, coordinated with the authorities, notified affected entities, and shared technical indicators with our partners to mitigate similar campaigns.

We believe that this group's abuse of Claude was able to substantially increase the speed and scale of the attack. Importantly, however, our takeaway is that this is not a story just about Claude, nor about what the attack was able to accomplish.

This challenge is not unique to Anthropic — every frontier model developer will face increasingly sophisticated attempts by threat actors to circumvent safeguards and misuse their models. What we observed here is one data point on a trendline. As models become more capable, we expect a wider swath of threat actors will continue to seek ways to misuse models for malicious ends. That is why the entire industry, along with government partners, must continue to strengthen our defenses.

Details of the CCP-Backed Cyber Espionage Campaign

The attackers developed a framework designed to execute components of their cyber espionage campaign in a way that relied on human input at a few key points but which was able to misuse Claude Code (a popular product of ours that enables Claude to autonomously write and execute code) and open standard Model Context Protocol (MCP) tools to execute many components of the cyberespionage campaign with a substantial degree of autonomy.[3] Using this combination of tools, the attackers circumvented our safeguards and deceived the model about the true nature of the tasks they were directing Claude to complete.

The campaign consisted of distinct phases. At first, a human operator input a target — for example, an entity, or an entity's network — to Claude. The framework's orchestration engine

[2] "Usage Policy." Anthropic, Sept. 15, 2025, https://www.anthropic.com/legal/aup
[3] "Introducing the Model Context Protocol." Anthropic, Nov. 24, 2024, https://www.anthropic.com/news/model-context-protocol

would then task Claude to autonomously conduct reconnaissance against multiple targets in parallel. Approximately 30 systems from foreign governments and global companies were targeted, consistent with the threat actor's instructions. Upon completion, Claude delivered results to the operators for review and to determine the next step.

Next, acting on the threat actor's direction, Claude leveraged third-party software tools to search for vulnerabilities in these systems. Claude looked for "weak spots" in the target's infrastructure that could be exploited for the operators to gain unauthorized access to these systems. Many of these software tools were the same open source software tools used by legitimate defensive actors.

The next and final step was to attempt to exploit any discovered vulnerabilities using third party tools and to then find and extract sensitive information. This was only successful in a handful of cases, but required similar abilities to scan for systems containing valuable information, identify and exploit vulnerabilities, and exfiltrate the information. It also involved "moving laterally" within the system to establish access to new areas of the target's system. At the threat actor's direction, Claude queried databases, extracted information, parsed results to identify proprietary information, and categorized findings by intelligence value to the human operator. Claude then produced a summary report for the human operators to review.

This attack demonstrated that current frontier AI models are capable of uplifting dedicated, sophisticated groups.[4] Our preliminary estimate is that the threat actor was able to leverage Claude to perform the work of a 10-person team managed by one human operator. For example, we observed that approximately 80 to 90% of the CCP-backed campaign tasks were automated by Claude, whereas the remaining 10 to 20% were tasks where the human operators reviewed Claude's outputs and directed the models.

There were critical limitations in the campaign. First, the models frequently hallucinated. Hallucinations are when models essentially "make up" incorrect information — in this case, false credentials, or that it had succeeded when in reality it had not. This means human operators have to spend more time carefully validating all claimed results, limiting overall operational effectiveness. Second, the attack still fundamentally required a human operator at various decision points to progress. That is, the models still requested approval to progress from reconnaissance to active exploitation, authorize use of harvested credentials, and to make final decisions about data exfiltration. Lastly, the campaign did not produce fundamentally novel attack techniques unknown to security practitioners. Rather, it applied existing methods to identify and exploit vulnerabilities in software systems at scale.

[4] "Uplift" is the term we use to estimate how much individuals are able to benefit from using models compared to if they had tried to accomplish the same outcome without using models.

Anthropic's Work to Disrupt the CCP-Backed Espionage Campaign

Anthropic detected this CCP-backed campaign within two weeks of the attackers' first confirmed offensive activity. Anthropic maintains multiple systems designed to detect suspicious activity, including cyber classifiers and what are known as YARA rules in the security industry.[5] In this case, one of these systems triggered an immediate human investigation. Over the following 10 days, we banned the associated accounts, implemented detection mechanisms for similar behavior, notified affected entities, and coordinated with authorities to gather actionable intelligence. We also collected the technical indicators of these attacks, and took steps to share these with partners, including other frontier labs, with whom we have threat sharing agreements, so that they could identify and mitigate similar campaigns.

We assessed with high confidence that the threat actor was affiliated with the CCP because of technical evidence from the sophisticated obfuscation infrastructure that enabled the threat actor to access Claude accounts and evade detection. In addition, the targeted entities aligned with known targets of the CCP; and the operators exhibited behavior consistent with this conclusion, including following the Chinese workday — including observing lunch breaks — and observing Chinese national holidays.

The threat actor went to great lengths to obfuscate their work, conceal their intentions from Claude, or evade our safeguards. First, the actor "jailbroke" our models by, in some instances, deceiving the model, falsely stating they were conducting ethical defensive cybersecurity testing. Then, having convinced the models to comply, the attackers created a sophisticated network of many accounts, which all used separate instances of the model to perform subcomponents of the attacks on different targets. Separating work in this way frequently makes the subcomponents seem benign, but when put together, form a pattern of malicious behavior. They routed their actions through an obfuscated network they controlled.

Anthropic is Continuing to Secure its Models in Response to this Campaign

During and after the campaign, we instituted new mitigations to better prevent this kind of misuse of Anthropic models. We expanded our detection mechanisms to better cover novel threats such as this campaign — including by improving our cyber-focused classifiers. We are also prototyping early detection systems specifically targeted at autonomous cyber attacks, and researching new techniques for investigating and mitigating large-scale distributed operations.

[5] "Using YARA For Malware Detection." NCCIC, https://www.cisa.gov/sites/default/files/FactSheets/NCCIC%20ICS_FactSheet_YARA_S508C.pdf

Importantly, because all AI models are susceptible to this type of misuse, we shared and continue to share the results of our investigation with frontier labs. Defensive actors worldwide need to prepare for and defend against these new threats.

What Industry and Government Should Do

As model capabilities advance, AI developers have to get better at understanding risks, preventing misuse, and ensuring that models can be used by defenders. This is a shared challenge on which industry and government should work together. While the threat actors likely leveraged Claude for this campaign due to its advanced coding and agentic capabilities, many models available today could soon be able to conduct such an attack. It is therefore critical that industry, government, and researchers work together to evaluate model capabilities, rapidly secure critical infrastructure, and develop better methods to restrict malicious use.

Predeployment Testing and Transparency for National Security Capabilities

The United States should continue to be the best and fastest at evaluating model capabilities, deploying models, and learning from these deployments. Government-led evaluations remain critical, as the Intelligence Community and agencies like the Department of Energy possess unique expertise to evaluate how adversaries could exploit AI models.

The Frontier Red Team has an ongoing partnership with the U.S. government that enables risk mitigation and provides strategic national security insights. One major part of this is our collaboration with the U.S. Center for AI Standards and Innovation (CAISI) in the Department of Commerce. Through voluntary agreements, the CAISI conducts rapid predeployment testing of our Claude models that gives the government visibility into AI model capabilities, provides us with critical information about our models' national security implications, and allows us to launch our commercial models more rapidly and with enhanced confidence about their reliability. Because of the sensitive nature of cybersecurity information, the CAISI and the U.S. government in general are in an advantageous position to evaluate model capabilities and understand capability trajectories better than anyone in the world. Codifying the CAISI can ensure the government can test and evaluate models for these capabilities, in partnership with the U.S. national security community.

In conjunction with government testing, transparency standards play a crucial role in achieving secure AI development. This is why Anthropic published a transparency framework to inform light-touch guardrails that encourage the largest AI developers to follow secure practices — disclosing how they assess and mitigate national security risks, their testing procedures, and

results.[6] This transparency approach would establish industry best practices for safety and set a baseline for secure model training, ensuring developers meet basic accountability standards while enabling public visibility into development without impeding innovation.

Threat Intelligence Sharing

Additionally, the U.S. government has an important role in identifying what critical national infrastructure must be protected. We know that all American frontier AI labs are targets for infiltration by state and non-state actors. As the models become more capable, it is critical that frontier labs work with the U.S. government to implement defensive measures against threat actors who would seek to abuse their models. This is why we believe there should be more robust channels between American frontier AI laboratories and the U.S. government to facilitate threat intelligence sharing, similar to information sharing processes used in critical infrastructure sectors, so we may shore up our collective defenses against malicious actors. Galvanizing the U.S. government and industry capacity to sprint to prepare AI infrastructure for a world of cybersecurity AI agents is critical at this juncture.

Making Models Useful for Cyber Defenders

We therefore think a large part of making the future secure depends on our ability to make models useful for defenders and get the models into those defenders' hands. To that end, Anthropic has piloted and deployed our models with a large fraction of the world's largest cybersecurity companies, with whom we continue to partner.

We are also developing tools designed to help defenders. For example, Anthropic has released a security review tool that, with a single command, reviews a codebase for vulnerabilities and can suggest patches before code reaches production.

We envision a world where models are used by cyberdefenders — in industry, government, and by individual researchers and engineers — to secure all parts of the infrastructure that the world relies on. I am particularly encouraged by a new generation of advanced startups that are among the fastest and best at deploying models in creative ways to outpace attackers. We believe it is very possible that the force of innovation, spearheaded by inventive white hat companies, will be the most important factor in our ability to triumph over threat actors.

The Stakes of Maintaining U.S. Leadership in AI

This campaign also underscores a broader strategic reality: the United States and like-minded

[6] "The Need for Transparency in Frontier AI." Anthropic, July 7, 2025, https://www.anthropic.com/news/the-need-for-transparency-in-frontier-ai

democracies must maintain leadership in frontier AI development. Based on the current trajectory of AI development, our ability to lead at the AI frontier in the 2026-2027 time period will likely also translate directly into significant capability advancements in cyber, military, intelligence, and other critical national and economic security functions.

In this case, CCP-sponsored operators misused an American model running on American infrastructure because our technology represents the state of the art. That's not a coincidence — it's a direct result of U.S. policy choices that have constrained the CCP's access to the advanced compute needed to train frontier models. Because CCP-sponsored operators had to use our systems, we were able to detect and disrupt them, and share information about the threat with the U.S. government. That is an enormous strategic advantage.

The Trump administration has already taken important steps to advance U.S. AI leadership, including accelerating the domestic buildout of AI infrastructure, promoting federal adoption, and strengthening safety testing and security coordination. But preserving the United States' lead in frontier AI development during this critical window depends on protecting our current advantage in compute — or the AI chips that power advanced AI systems. Restrictions on exports of advanced semiconductors and semiconductor manufacturing equipment to the CCP, building on actions initiated during the first Trump Administration and expanded under the Biden Administration, have been vital to preserving that edge.

Relaxing controls on advanced AI chips at this juncture could allow the CCP to close the gap in frontier AI development — producing models that may match or exceed current U.S. capabilities for cyber-offensive tasks, but without our safeguards, and using them to target U.S. critical infrastructure and national champions. Export controls on advanced semiconductors have proven effective at constraining the CCP's AI development. Without them, what any individual American company does to secure its own models becomes far less consequential. We simply won't see the attacks coming.

Conclusion

We are in a race against threat actors to secure systems faster and more robustly than they can be attacked. Threat actors will stop at nothing to develop, steal, or manipulate AI models to conduct increasingly sophisticated cyberattacks at scale, and we must respond urgently.

Thank you for the opportunity to appear before the Committee today, and I look forward to answering your questions.