Testimony of Royal Hansen,
Vice President of Privacy, Safety, and Security Engineering
U.S. House Committee on Homeland Security
Subcommittee on Cybersecurity and Infrastructure Protection &
Subcommittee on Oversight, Investigations, and Accountability
December 17, 2025

Chairmen Garbarino, Ogles, Brecheen; Ranking Members Thompson, Swalwell, Thanedar; and Members of the Committee and Subcommittees: thank you for the opportunity to speak with you today. My name is Royal Hansen, and I serve as Vice President of Privacy, Safety, and Security Engineering at Google. Our team is responsible for building and scaling the foundational technology to keep billions of people safe online.

Thank you for holding this important hearing. We welcome the opportunity to provide information about Google's efforts to secure its own artificial intelligence, protect its customers' workloads, and use artificial intelligence to strengthen cyberdefense and enhance our collective security.

Securing our Artificial Intelligence

Google's AI principles, published in 2018 and updated this year, describe our commitment to developing technology responsibly and in a manner that is built for safety, enables accountability and upholds high standards of scientific excellence. We have built on this work through our Secure AI Framework, as well as with extensive model hardening and various governance measures. This comprehensive approach means we secure all components of the AI ecosystem including data, infrastructure, applications, and models.

The Secure AI Framework (SAIF)

SAIF is our framework for integrating security and privacy measures into machine learning and generative AI applications and it governs how we embed controls throughout the AI system stack from data, infrastructure, application and models. The framework, which is designed to ensure that AI models are secure by design, has six core elements:

- **Expand strong security foundations to the AI ecosystem.** Leverage secure-by-default infrastructure protections and expertise built over the last two decades to protect AI systems, applications and users. At the same time, develop organizational expertise to keep pace with advances in AI and start to scale and adapt infrastructure protections in the context of AI and evolving threat models. For example,

injection techniques like SQL injection have existed for some time, and organizations can adapt mitigations, such as input sanitization and limiting, to help better defend against prompt injection style attacks.

- **Extend detection and response to bring AI into an organization's threat universe.** Detect and respond to evolving AI-related cyber incidents by extending threat intelligence and other capabilities. For organizations, this includes monitoring inputs and outputs of AI systems to detect anomalies, and using threat intelligence to anticipate attacks. This effort typically requires collaboration with trust and safety, threat intelligence, and counter abuse teams.

- **Automate defenses to keep pace with existing and new threats** Harness the latest AI innovations to improve the scale and speed of response efforts to security incidents. Adversaries will use AI to scale their impact, so it is important to use AI and its current and emerging capabilities to stay nimble and cost effective in protecting against them. It is important to remember that the vast majority of successful attacks - whether AI-enabled or not- prey on legacy systems; AI can help defenders modernize and address issues at a scale and speed that has historically proved challenging.

- **Harmonize platform level controls to ensure consistent security across the organization.** Align control frameworks to support AI risk mitigation and scale protections across different platforms and tools to ensure that the best protections are available to all AI applications in a scalable and cost efficient manner. (could we say something like - we can't have independent defense structures for AI and legacy systems as attackers prey on the gaps between such approaches) At Google, this includes extending secure-by-default protections to AI platforms like Vertex AI and Security AI Workbench, and building controls and protections into the software development lifecycle. Capabilities that address general use cases, like Perspective API, can help the entire organization benefit from state of the art protections.

- **Adapt controls to adjust mitigations and create faster feedback loops for AI deployment.** Constantly test implementations through continuous learning and evolve detection and protections to address the changing threat environment. This includes techniques like reinforcement learning based on incidents and user feedback, and involves steps such as updating training data sets, fine-tuning models to respond strategically to attacks, and allowing the software that is used to build models to embed further security in context (e.g. detecting anomalous behavior). Organizations can also conduct regular Red Team exercises to improve safety assurance for AI-powered products and capabilities. These are exactly the techniques we have used to defend Gmail, the Play Store and Chrome with AI at scale for many years.

- **Contextualize AI system risks in surrounding business processes.** Conduct end-to-end risk assessments related to how organizations will deploy AI. This includes an assessment of the end-to-end business risk, such as data lineage, validation and operational behavior monitoring for certain types of applications. In addition, organizations should construct automated checks to validate AI performance. Nearly all businesses are increasingly digital - AI will only accelerate that trend. The controls required to mitigate risks in these processes must keep pace - some of which will be digital and some will be procedural.

Model Hardening

Our AI models are fine-tuned on large datasets of realistic attack scenarios to build intrinsic resilience. They are taught to recognize and ignore malicious instructions while still following user requests. This is, and will continue to be, an evolving space requiring rapid iterations as attackers innovate.

Over the past decade, we have evolved our approach to translate the concept of red teaming to the latest innovations in technology, including AI. The AI Red Team is closely aligned with traditional red teams, but also has the necessary AI subject matter expertise to carry out complex technical attacks on AI systems. A core part of our security strategy is automated red teaming, where our internal Gemini team constantly attacks Gemini in realistic ways to uncover potential security weaknesses in the model. We fine-tuned Gemini on a large dataset of realistic scenarios, where automated red teaming generates effective indirect prompt injections targeting sensitive information.

Protecting AI models against attacks like indirect prompt injections requires “defense-in-depth” – using multiple layers of protection, including model hardening, input and output checks (like classifiers), and system-level guardrails. Securing advanced AI systems against specific, evolving threats like indirect prompt injection is an ongoing process. It demands pursuing continuous and adaptive evaluation, improving existing defenses and exploring new ones, and building inherent resilience into the models themselves.

Securing AI Workloads

Recent headlines have highlighted several key vulnerabilities and attack vectors targeting private and public sector entities. It is clear that legacy systems, misconfigured cloud environments, and the exploitation of known vulnerabilities remain significant concerns. Email phishing, supply chain attacks, criminal hacking, and state-sponsored cyber espionage further compound these challenges. Our approach to protecting public and private sector entities is built on several core tenets:

- AI-Powered Security: We leverage the power of AI and machine learning to enhance threat detection, automate security operations, and secure AI development.
- Secure by Design: We engineer security into every layer of our infrastructure and services, from custom-designed hardware to advanced encryption techniques. To do this well requires security engineering which goes well beyond checklists and compliance requirements.
- Zero Trust: We ensure that no user or device is inherently trusted, regardless of their location or network. Access is continuously authenticated and authorized based on identity, device health, and context. We developed this approach in the wake of Chinese threat actor attacks on Google over 15 years ago, and it remains as important today.
- Shared Fate: We operate under a clear shared responsibility model, securing the underlying cloud infrastructure while providing tools and guidance for customers to manage their own security. We believe in a "shared fate" where our success is tied to the customer's. We are deeply invested in the collective security outcomes of consumers, companies and countries. We align our goals with the security and resilience of critical operations, particularly where national security is at stake.

Artificial Intelligence and Cybersecurity: Identifying Opportunities and Mitigating Risks

We stand at a critical technological inflection point. Rapid advances in AI are unlocking new possibilities for the way we work and accelerating innovation in science, technology, and beyond. Some of these same AI capabilities, however, can also be deployed by attackers, leading to understandable anxieties about the potential for AI to be misused for malicious purposes. Until recently, our analysis of government-backed threat actor use of AI revealed that threat actors were using generative AI primarily for common tasks like troubleshooting, research, and content generation. Over the past year, Google Threat Intelligence Group has identified an important shift, with adversaries not only leveraging AI for productivity gains, but experimenting with novel AI-enabled malware in active operations.

We have identified malware families that use LLMs to generate malicious scripts, obfuscate their own code to evade detection, and use AI models to create malicious functions on demand, rather than hard-coding them into the malware. This marks a new operational phase of AI abuse, involving tools that dynamically alter behavior mid-execution. While still nascent, this development represents a significant step toward more autonomous and adaptive malware. We have and will continue to publish on these topics, take action and enhance our products to ensure industries and societies as a whole can keep pace with the latest threats.

Today, and for decades, the main challenge in cybersecurity has been that attackers need just one successful, novel threat to break through the best defenses. Defenders, meanwhile, need to deploy the best defenses at all times, across increasingly complex digital terrain — and there is no margin for error. As we have seen in recent years, this is particularly true for legacy technology. This is the “Defender's Dilemma,” and there has never been a reliable way to tip that balance.

Our experience deploying AI at scale informs our belief that AI can reverse this dynamic in several ways and enhance our collective security.

- AI allows security professionals and defenders to scale and accelerate their work in threat detection, malware analysis, vulnerability detection, vulnerability fixing and incident response.

- Google’s AI-based efforts like BigSleep have demonstrated AI’s ability to find new zero-day vulnerabilities in well-tested, widely used software. Developed by Google DeepMind and Google Project Zero, Big Sleep can help security researchers find zero-day (previously-unknown) software security vulnerabilities. Since it was introduced last year, it has continued to discover multiple flaws in widely-used software, exceeding our expectations and accelerating AI-powered vulnerability research. With Big Sleep, we have demonstrated how we can find vulnerabilities that defenders don’t yet know about. In this case, we found a vulnerability that the attackers knew about and had every intention of using. We were able to detect and report it for patching before they could exploit it.

- Finding vulnerabilities is only half of the battle. Recently, we developed CodeMender, an AI-powered agent that utilizes the advanced reasoning capabilities of our Gemini models to automatically fix critical code vulnerabilities. CodeMender scales security, accelerating time-to-patch across the open-source landscape. It represents a major leap in proactive AI-powered defense and includes features such as root cause analysis and self-validated patching. This capability in particular will be the most significant security advancement in many years.

Collaboration Toward Responsible Artificial Intelligence Adoption

We believe the private sector, governments, educational institutions, and other stakeholders must work together to maximize AI's benefits while also reducing the risks of abuse. As innovation moves forward, the industry more broadly needs security standards for building and deploying AI responsibly. That's why Google introduced SAIF, as noted above, as a conceptual framework to secure AI systems. Our recent expansion to SAIF 2.0 addresses the rapidly

emerging risks posed by autonomous AI agents and extends our proven framework with new guidance on agent security risks and controls to mitigate them.

In addition, Google co-founded the Coalition for Secure AI (CoSAI), an open-source initiative to help all developers and deployers of AI create and maintain secure by design AI systems and help advance the framework. CoSAI helps foster a collaborative ecosystem to share open-source methodologies, standardized frameworks, and tools. Since its launch, CoSAI has made significant strides in strengthening AI security in collaboration with industry and academia in areas including Software Supply Chain Security for AI Systems; Preparing Defenders for a Changing Security Landscape; AI Security Risk Governance; and Secure Design Patterns for Agentic Systems. We have also supported the MLCommons Association's efforts to develop AI safety benchmarks by contributing funding for the development of a testing platform, as well as technical expertise and resources. ML Commons' shared research infrastructure helps the scientific research community derive new insights for breakthroughs in AI.

Across Google Cloud, we model and promote the adoption of responsible AI data practices that preserve our customers' privacy and support their compliance journey. Robust privacy commitments outline how we protect user data and prioritize privacy and the greater adoption of artificial intelligence rearms their importance. We adhere to a holistic approach to AI risk management and compliance, including focusing on employing an AI risk assessment methodology for identifying, assessing, and mitigating risks; developing and using an automated, scalable, and evidence-based approach for auditing generative AI workloads; and emphasizing human oversight and collaboration in our risk assessments and governance councils.

We use explainability tools to help understand and interpret AI predictions and evaluate potential bias; privacy-preserving technologies such as masking and tokenization and adhering to privacy laws; continuous monitoring and auditing for security vulnerabilities that AI might miss; investing in training programs to bridge the AI knowledge gap; and encouraging "interdisciplinary collaboration" between data scientists, risk analysts, and domain experts is also key.

Cybersecurity has never been a field where perfection is possible. It will remain a dynamic space for years to come, and speed and resilience will be required to defeat and contain innovative attackers. As governments and civil society leaders look to counter evolving threats from cybercriminals and state-backed attackers, we are committed to leading the way in using AI to tip the balance of cybersecurity in favor of defenders.

We appreciate the Committee convening this important hearing. And we look forward to answering your questions.

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.

House Homeland Security: **“The Quantum, AI, and Cloud Landscape: Examining Opportunities, Vulnerabilities, and the Future of Cybersecurity”**

December 17, 2025

Good morning, Chairman Garbarino, Ranking Member Thompson, Chairman Ogles, Chairman Brecheen and members of the committee. Thank you very much for the opportunity to testify today.

My name is Eddy Zervigon, and I am the CEO of Quantum XChange. We were founded in 2018, two years after NIST was tasked with evaluating the algorithms to take us into the quantum age. Quantum XChange is a cybersecurity company that interoperates with the major network infrastructure vendors to enable the encryption that protects data today and into the post-quantum future with hardware and software solutions developed entirely in the United States.

While quantum computing and AI promise new breakthrough capabilities, they also introduce significant risks to our national and economic security that must be urgently addressed. AI can enable faster, more dangerous cyberattacks and quantum computers can break current encryption standards, exposing sensitive data. These capabilities will be weaponized by our adversaries, creating a very dangerous imbalance in our cyber defenses.

For more than 50 years, encryption has safeguarded our data from theft and misuse. We’ve had the luxury of a “set it and forget it” mindset, trusting its strength by default. That era is now ending with quantum computing.

Think about it like this:

Imagine all digital communications from government agencies sent over the past ten years being readable by our adversaries. This is a real threat to the US today; rogue nation states and state-sponsored terrorist groups are collecting encrypted data NOW to decrypt later with a quantum computer.

Further, now imagine our adversaries reading sensitive government data in real time, and altering it without anyone knowing. This could be tomorrow’s reality.

Public and private sector work on quantum-resilient solutions is ongoing. Technologies, like post-quantum cryptography (PQC) *or quantum-safe encryption algorithms*, are part of the solution but not the complete answer. Despite our best efforts, post-quantum cryptography may still be vulnerable to quantum-enabled attacks.

All of which raises this fundamental question and challenge:

> *What happens when an algorithm breaks (because it is a when, not if)? Every agency CIO, enterprise CISO, security vendor, and network gear manufacturer must be able to answer that question.*

In our view, what’s needed to ensure data security and confidentiality in the quantum age is an architectural approach, not just a new algorithm.

This architectural approach enables agencies to focus on securing the network that data travels on to strengthen the existing infrastructure against quantum attacks, while minimizing disruption to existing operations. This is how our government agencies need to be protected. When you have valuables in your house, the first step isn’t buying a new jewelry box with biometric access controls, it's locking your front and back doors, so the house is secure and harder to get in. Once your home is secure, then you can figure out what specific rooms need further locks or security measures to protect your valuables and sensitive documents.

Federal agencies handling sensitive data need to act now and follow the lead set by Customs and Border Protection. Our work with CBP to incorporate PQCs across their network infrastructure in 2026 has shown that you **can** begin to secure your networks *today* with quantum-resistant technologies in a FIPS validated way, without having to rip and replace your entire infrastructure. I cannot stress enough that timing here is critical.

Agencies that fail to prepare today risk leaving their data vulnerable. Every day that we are not quantum-resistant is another day that data is harvested, to be decrypted later. It is important to note, that we at Quantum XChange are not the only ones advocating for action today. The Quantum Industry Coalition, which we are a part of and includes Amazon Web Services, Google, IBM, Microsoft, Accenture, and others believes “that *agencies handling sensitive government data should already be actively preparing for the transition and should begin migrating high-risk systems to FIPS/NIST validated PQC where possible.*

Having the opportunity to meet with several of your offices, I was often asked “What can Congress do?” Through this committee’s leadership, and building off the work previously done, Congress can accelerate the timelines for PQC compliance, allocate the budget to allow the migration process to begin, and work with leaders within the Administration to encourage adoption, as the technology is readily available and deployable today. America’s defenses cannot stop at our physical borders. Through your leadership and efforts, and in partnership with private sector partners like us, we can and will secure America’s digital borders too.

In closing, I want to thank you all again for the opportunity to offer some thoughts today and look forward to your questions.

Eddy Zervigon

Quantum Industry Coalition Position on Post-Quantum Cryptography

October 23, 2025

The National Institute of Standards and Technology (NIST) has approved the first set of post-quantum cryptographic (PQC) algorithms, in what promises to be an iterative process moving forward. NIST has been leading the migration charge for close to a decade, evaluating and approving the algorithms and delivery architectures that will protect our data networks into the post-quantum era.

The Federal government has set timelines for the adoption of these post-quantum algorithms through legislation and executive orders. Government agencies should already be preparing for PQC transition through education, cryptographic inventory, risk assessments, transition strategies, and pilots. At the same time, the ecosystem of innovative start-ups and established players surrounding the delivery of these algorithms has progressed to a point where transition is possible in some high-risk areas, such as securing the network layer.

It is our position that agencies handling sensitive government data should already be actively preparing for the transition and should begin migrating high-risk systems to FIPS/NIST validated PQC where possible.

Quantum Industry Coalition Members Include:

Accenture
D-Wave
Entanglement Institute
IonQ
Quantinuum
Rigetti Computing
Xanadu
Amazon Web Services
Cold Quanta
Diraq
Google
MesaQuantum
Quantum Corridor
SandboxAQ
Anametric
EeroQ
IBM
Microsoft
Quantum Machines
SEEQC
Atom Computing
enQase
Infleqtion
Qolab
Quantum XChange
Strangeworks

House Homeland Security Committee

Subcommittee on Cybersecurity and Infrastructure Protection
Subcommittee on Oversight, Investigations, and Accountability

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

Written Testimony of Michael Coates
Founding Partner, Seven Hill Ventures

December 17, 2025

Chairman Ogles, Ranking Member Swalwell, Chairman Brecheen, and Ranking Member Thanedar, I thank you for the opportunity to testify before you today. I'm honored to be here to speak about the changing landscape in cybersecurity and the resulting impacts from AI and quantum computing.

The perspective I will share is grounded in over twenty years of experience in cybersecurity, including service as a chief information security officer, a chairman of a global non-profit advancing the state of application and coding security, a technology startup founder, and a venture capital investor supporting cybersecurity innovation.

Today we sit at the precipice of significant change. While advancements in AI and development towards AGI are widely discussed, the practical and operational impacts to cybersecurity defenders are less often examined.

The fundamental reality is not that AI and quantum are creating new types of threats, but rather they are collapsing the time, cost and skill required to conduct cyber operations. These changes are outpacing the existing technical, regulatory and operational defenses. This shift reshapes the cyber threat landscape and forces a reconsideration of how we defend critical systems in an era defined by speed, automation, and intelligent scale.

What Is Changing: The Compression of Cyber Capability

Capability Compression & Orchestration Expands the Attacker Base

Corporations and citizens potentially face a variety of threat agents including highly funded nation state adversaries, financially motivated cybercriminal organizations, and lone hacktivists motivated by ideology. Each attacker type has different skills and resources at their disposal and to date, these have constrained the complexity or scale of cyberattacks available to each adversary.

The most advanced attacks were often only launched by nation state adversaries against select targets. Whereas cybercriminal entities focused their efforts on pipelines of optimized offensive security services, such as ransomware extortion, to monetize the compromise of businesses or individuals.

Robust security attacks require a series of steps spanning reconnaissance, exploitation, command and control, and delivery of the ultimate objective, such as data theft or system modification. Each of these components could be executed by a well-funded nation state adversary or a competent cybercriminal organization, but it was not as

achievable for the lone hacktivist or unsophisticated security hacker. This is rapidly changing.

As demonstrated in the November, 2025 Anthropic report “Disrupting the first reported AI-orchestrated cyber espionage campaign”[1], a nation state adversary used AI systems as a central brain and point of coordination for a complete security attack against multiple targets across the United States. AI was used to execute and interpret results for each step of the attack and as an overall orchestration layer, with the human adversary only interacting at a few decision points.
While this attack may not have demonstrated new or novel attack methods, the orchestration and use of AI is a critical development in the ecosystem of the cybersecurity adversary.

Agentic Attacks Remove Human Constraints

Agentic AI systems will enable the attacker to no longer be bound by time of day, hours awake, or the need for food or sleep. Autonomous agentic systems are replicating the most advanced attackers and will be able to target with accuracy and ease.
This is no longer theoretical as research just released by Stanford[2] shows that an autonomous AI penetration-testing agent already performs at or above the level of most highly skilled professional security testers, outperforming nine out of ten participants in a live network test with an 82% valid vulnerability discovery rate at a fraction of the cost.

Acceleration of Vulnerability Discovery and Exploitation

Furthermore, the increasing power of AI for software vulnerability analysis is enabling faster and more accurate detection of previously unknown zero day security vulnerabilities. For example, Google’s Big Sleep, a collaboration between Google Project Zero and Google DeepMind, has discovered a critical zero day vulnerability in the major software SQLite Database Engine.[3]

Over the past decades, the challenge for many organizations has not been knowledge that a vulnerability existed, but rather the operational inertia to deploy, test, and productize the software patch. In fact, the 2025 Verizon Data Breach Investigations Report found that vulnerability exploitation was the initial access vector in 20% of breaches, and that defenders often cannot remediate fast enough—organizations fully

[1] https://www.anthropic.com/news/disrupting-AI-espionage
[2] https://arxiv.org/pdf/2512.09882
[3] https://cloud.google.com/blog/products/identity-security/cloud-ciso-perspectives-our-big-sleep-agent-makes-big-leap

remediated only about 54% of vulnerabilities in network edge devices, with a median remediation time of 32 days, while CISA KEV vulnerabilities can be mass exploited in a median of five days.[4]

The Practical Result: Reduced Time for Defenders

With AI orchestration, the ease of launching comprehensive cybersecurity attacks against any target is substantially reduced. The result is that many more potential adversaries now have the means to execute these attacks.

In addition to an increase in attacks against the most critical targets, this development will also result in lesser-profile targets, such as small businesses across the country, being subjected to full-scale security assaults.

The direct result of this change will be a dramatic drop in the time available for defenders to detect attacks, initial compromise, or lateral movement before critical access or sensitive data is breached. Taken together, these shifts do not just increase cyber risk, they fundamentally change the speed at which cyber incidents unfold.

Why Time Compression Changes the Nature of Cyber Risk

The compression of time, cost, and skill required to conduct cyber operations fundamentally changes how cyber risk manifests in practice. While individual techniques may appear familiar, the speed at which attacks now unfold alters the balance between attackers and defenders in ways that existing security models were not designed to accommodate.

The most immediate consequence is a dramatic reduction in the time available for defenders to detect and respond to malicious activity. AI-enabled orchestration and automation allow attackers to move from initial access to lateral movement and impact far more quickly than in the past. In many cases, defenders are no longer responding to early indicators of compromise, but to attacks that are already well underway.

This compression of time disproportionately affects organizations that lack large, specialized security teams. While highly resourced enterprises may be able to invest in advanced detection and response capabilities, smaller organizations, including hospitals, schools, food processing facilities, and small businesses often rely on delayed or manual processes. As sophisticated attacks become easier to launch and

[4] https://www.verizon.com/business/resources/reports/dbir/

less expensive to operate, these lower-profile targets increasingly face the same level of adversarial capability once reserved for critical national infrastructure.

At the same time, intelligent automation and scaling by adversaries is shifting the risk of attacks from periodic events to a continuous threat. AI-driven attacks do not require sustained human attention and can operate persistently, adapting to defenses and retrying failed approaches automatically. This erodes traditional assumptions that organizations can recover between incidents or rely on periodic assessments to maintain security.

Existing defensive and governance models further compound this challenge. Over the past decades, many major breaches did not occur because vulnerabilities were unknown, but because organizations were unable to deploy patches or mitigations quickly enough. As AI accelerates vulnerability discovery and exploitation, this operational inertia becomes more consequential. The gap between awareness and action grows more dangerous as attack timelines compress.

The result is a widening gap between the speed and accessibility of modern cyberattacks and the ability of most organizations to respond. As AI compresses attack timelines and expands the pool of capable adversaries, cybersecurity outcomes will increasingly be determined by whether defenses can operate at machine speed.

Implications for Cyber Defense, Policy and Coordination

The advancements in artificial intelligence and quantum computing present significant opportunities for innovation, but without appropriate alignment between technology, operations, and governance, they also introduce material cybersecurity risk. The shifts described earlier are not theoretical, and they cannot be addressed by any single organization or sector acting alone.

The following are key areas where attention is warranted to increase the cybersecurity posture of our organizations and critical systems.

- **Secure by Design as a Baseline Expectation**

  As software is increasingly written, analyzed, and modified by AI systems, secure design principles must be integrated into the creation of software from the outset. Initiatives such as CISA's Secure by Design program, along with industry standards promoted by organizations like OWASP and the Cloud Security Alliance, provide important guidance. Supporting these organizations and

reinforcing these efforts helps ensure that speed and automation do not come at the expense of security fundamentals.

- **Regulatory Clarity That Supports Speed and Innovation**

Clear and transparent regulatory frameworks are necessary to enable rapid innovation while maintaining responsibility for security and safety. In an environment where threats evolve quickly, ambiguity or fragmentation in regulation can unintentionally slow defensive response and increase systemic risk. Policy should seek to provide clarity and consistency without constraining the ability of organizations to adapt at machine speed.

- **Public–Private Coordination on AI-Driven Cyber Threats**

The pace of change in the cyber threat landscape reinforces the importance of strong public–private partnerships. Effective coordination, information sharing, and joint response mechanisms help ensure that defensive learning keeps pace with adversarial innovation. These partnerships remain a critical component of national cyber resilience as AI-driven threats continue to evolve.

- **Migration Toward Autonomous Defensive Capabilities**

As attackers increasingly rely on automation and agentic systems, purely human-driven defenses will struggle to keep pace. Continued investment in research, development, and deployment of intelligent and autonomous defensive systems is necessary to address machine-speed threats. This includes supporting innovation across both the public and private sectors.

- **Quantum Preparedness for Cryptographic Systems**

Stable, cryptographically relevant quantum computing would render many of today's widely deployed public-key encryption algorithms ineffective, impacting secure communications across government, industry, and critical infrastructure. While post-quantum cryptographic standards already exist, the primary challenge is the time and coordination required to migrate existing systems. Deliberate preparation is crucial to avoid a reality where an adversary achieves cryptographically relevant quantum capabilities first and thus access not only to future communications, but potentially to sensitive data captured and stored today.

- **Trustworthiness and Transparency in AI Systems**

  As AI systems are increasingly embedded into security-sensitive workflows, trust in operation becomes crucial. Large language models reflect the data, incentives, and governance structures under which they are trained, and these factors can materially influence reliability and security outcomes.

  Bias in AI systems — whether intentional or unintentional — can affect how software is generated, how alerts are prioritized, and how decisions are made. In security-critical contexts, performance alone is not sufficient; the provenance, training, and oversight of AI systems must also be considered as part of risk assessment.

  Furthermore, greater transparency in software procurement and composition is needed. Requiring bill of materials and software contracts to disclose the use of AI within software, as well as the specific models and model origins, can help organizations better assess risk and make informed security decisions, particularly in sensitive or critical environments.

Artificial intelligence and quantum computing are accelerating dynamics that dramatically shift the cybersecurity landscape. As AI and quantum computing continue to advance and are increasingly leveraged by cyber adversaries, success will depend on whether our technical, operational, and institutional responses can adapt at comparable pace.

I appreciate the opportunity to share these observations and look forward to your questions.