Testimony of Sandra Joyce
VP, Google Threat Intelligence
Hearing on "The AI Security Landscape: How Frontier Models, Agentic AI, and AI Coding Tools Are Reshaping Cybersecurity and Critical Infrastructure Resilience"
House Committee on Homeland Security, Subcommittee on Cybersecurity and Infrastructure Protection
June 4, 2026

Chairmen Garbarino, Ogles; Ranking Members Thompson, Ramirez; and Members of the Committee and Subcommittees: thank you for the opportunity to speak with you today. My name is Sandra Joyce, and I serve as Vice President of Google Threat Intelligence Group (GTIG). Our team relentlessly defends Google, our users, and customers by building the most complete threat picture to disrupt adversaries.

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

Artificial Intelligence and Cybersecurity: Identifying Opportunities and Mitigating Risks

We stand at a critical technological inflection point. Rapid advances in artificial intelligence (AI) are unlocking new possibilities for the way we work and accelerating innovation in science, technology, and beyond. This technology has impacted cybersecurity in profound ways for the defender as well as the attacker.

While recent introductions to AI vulnerability research have caught the attention of mainstream audiences, Google has long anticipated and prepared for this impact. Over the years, our security teams and frontier AI research teams at Google DeepMind have introduced tools like CodeMender, which have shown the immense potential of AI to discover and mitigate vulnerabilities. For years we have used these tools to review and harden the open source software that many of us depend on.

While Google leveraged AI for defense, we anticipated that threat actors would abuse these capabilities to find and exploit vulnerabilities.[1] Those concerns were validated recently when we

[1] Google Threat Intelligence Blog | Google Cloud

discovered a criminal actor had used AI to develop a zero-day exploit.[2] We expect threat actors to continue attempting to use this technology to their advantage.

Over the past year, GTIG has identified an important shift, with adversaries not only leveraging AI for productivity gains, but successfully adopting AI to significantly enhance the scale, speed, and sophistication of their operations.

Looking ahead, we are particularly concerned about two future scenarios:

- **A deluge of vulnerabilities:** Though AI will be used by defenders to harden software and produce safer code, adversaries may have the initiative in the short term to find and exploit vulnerabilities at scale.
- **A swifter, scaled adversary that will overwhelm contemporary security:** Agentic orchestration allows threat actors to cheaply scale their operations and operate at unprecedented speed to take advantage of slow patch cycles, beleaguered security teams, and human response time.

An Evolution in Vulnerability and Exploitation

Along with Google, several of our peers have validated the potential of this technology to perform vulnerability and exploit development. Because exploits are the top attack vector in intrusions we observe,[3] this evolution has enormous consequences.

Though newer AI models can be used to find these vulnerabilities, less sophisticated models have been maliciously used for this purpose when coupled with a purpose-built harness, which is software designed to wrap around the model and make it operational. Adversaries do not require unprecedented breakthrough model capabilities; instead, they are deploying sophisticated harnesses to achieve an automated research and development capability.

Recently, GTIG discovered a campaign where cybercriminals utilized an AI model to support the discovery and weaponization of a zero-day vulnerability.[4] The actor planned to leverage the zero-day in a mass exploitation scheme. Google worked directly with the impacted vendor to responsibly disclose and patch this vulnerability before mass exploitation could occur.

Though this criminal example is the only confirmed development of a zero-day exploit using AI by a threat actor that we have thus far observed, we can infer that states have been doing

[2] Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access | Google Cloud Blog
[3] M-Trends 2026 Executive Edition
[4] Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access | Google Cloud Blog

significant research in this area and we expect that other zero-days developed with AI are already in use.

We are working to integrate AI directly into the development cycle and make code more difficult to exploit than ever; however, this transition period presents challenges. As we harden existing software with AI, threat actors will simultaneously use it to discover and exploit novel vulnerabilities.

The Agentic Shift

The most critical structural shift observed over the last year is the shift from users accessing LLMs through manual prompts to users deploying agents that can operate autonomously. Just as businesses are exploring the capabilities of agentic AI to automate their workflows, threat actors are experimenting with agentic capabilities to automate malicious activity, such as persistently probing a target or carrying out research and development on their behalf.

We recently observed a suspected PRC-nexus threat actor deploying agentic capabilities against Asian tech firms.[5] By incorporating open source tools and a memory system, the agent could map the target and autonomously pivot between tools based on its internal reasoning. Simultaneously, the actor leveraged a multi-agent penetration testing framework to automate the identification and validation of vulnerabilities. This approach suggests a transition toward autonomous reconnaissance that can scale the probing of targets with minimal human oversight.[6]

In another incident, we observed the North Korean actor, APT45, sending thousands of repetitive prompts to recursively analyze exploits.[7] This resulted in a more robust arsenal of exploit capabilities that would be impractical to manage without AI assistance.

In addition to the scale advantage conferred by agentic capabilities, threat actors are able to move rapidly before and after gaining access to a network using AI. Threat actors are using AI to take advantage of recently disclosed vulnerabilities before patches can be applied. After gaining access, agentic attacks can move rapidly through a network. In both cases, industry standards are based on human response time and insufficient to mitigate the threat.

---

[5] Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access | Google Cloud Blog
[6] GTIG AI Threat Tracker: Distillation, Experimentation, and (Continued) Integration of AI for Adversarial Use | Google Cloud Blog
[7] Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access | Google Cloud Blog

Defending the AI Supply Chain

As organizations continue integrating LLMs into production environments, the AI software ecosystem has emerged as a primary target for exploitation. While frontier models themselves remain highly resilient to direct compromise, adversaries are deploying traditional supply chain tactics against the orchestration layers—including open-source wrapper libraries, API connectors, and configuration tools.

The implications of this vector were demonstrated by a prominent cybercriminal cluster tracked by Google as UNC6780 (also known publicly as TeamPCP). The actor claimed responsibility for multiple supply chain compromises of popular GitHub repositories and associated GitHub Actions.[8] The compromise highlights the expanding attack surface of AI platforms and the potential for impact across the software supply chain. Given the package's widespread use, this incident could lead to considerable exposure of AI API secrets from affected victims, which could be used to gain further access to systems for traditional intrusion operations.

To help mitigate these supply-chain risks, we recommend that AI agents integrate automated security scanning directly into their public skill marketplaces. Every skill published to the repository should be analyzed to detect unauthorized network operations, malicious payloads, or unsafe embedded instructions. Based on this security-focused analysis, skills should be either approved as benign, flagged with user warnings, or blocked entirely, providing an essential layer of defense against ecosystem abuse.

Protecting the integrity of this supply chain cannot be treated as a standard, retroactive patch-management exercise; it requires a coordinated, structural shift in governance and defense. To secure these critical building blocks, organizations must establish comprehensive incorporation of AI capabilities into software supply chain security practices, including efforts to ensure transparency into model lineage, training datasets, and orchestration components. Organizations must also enforce strict least-privilege, zero-trust guardrails around AI data pipelines, ensuring autonomous agents operate in segmented environments without authority to elevate permissions or communicate with untrusted networks.

Autonomous Cyber Defense: Transitioning to Continuous, Lifecycle Security

Finding a vulnerability is only a single element of a broader operational challenge; to effectively tilt the cybersecurity balance in favor of defenders, we must close the exploit window entirely. Historically, patch management has been a retroactive, human-paced race against adversaries.

[8] Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access | Google Cloud Blog

In the current threat landscape, where attackers use AI to discover and target design flaws at scale, traditional, siloed security tools fail to keep pace.[9] Our threat intelligence demonstrates that modern digital risk is no longer confined to isolated software code errors. Real-world attack paths routinely emerge from the complex, real-time friction between cloud application interfaces, infrastructure configurations, permissions, and network identities. If a security team is handed an unprioritized list of thousands of software flaws, they face immediate patch fatigue. For critical infrastructure operators and public sector networks, defense-at-scale requires an automated mechanism that shifts focus away from mere bug hunting and toward comprehensive environmental exposure management.

The Operational Framework: An Autonomous Defensive Control Loop

To address this collapse of the exploitation timeline, Google has pioneered an always-on four-step framework designed to help enterprises to implement an autonomous defensive control loop: Prepare, Scan & Prioritize, Remediate, and Monitor.[10] Our aim is to shift the industry away from reactive response and toward active prediction and accelerated remediation.

- **Proactive Attack Surface Reduction (Prepare):** Before a vulnerability is ever targeted, defensive systems must establish a real-time exposure map of an entire network. By utilizing AI-driven, context-aware simulation agents, defenders can continuously pressure-test their own infrastructure. These agents behave exactly like an active adversary, mapping out complex, multi-stage attack paths to discover whether a sensitive asset is actually reachable from untrusted external networks. If a flaw cannot be reached, its immediate risk drops, allowing organizations to aggressively shrink their actual, real-world attack surface without relying on manual triage.

- **Contextual Risk Validation (Scan & Prioritize):** When new vulnerabilities are surfaced by AI code security agents—such as CodeMender, which leverages large language models to locate dormant, pre-authentication flaws—they must be cross-referenced with live operational data. By merging deep threat intelligence with cloud-posture data, AI-driven defense identifies which vulnerabilities represent an existential threat to critical workflows, sorting out benign software anomalies from active crisis points.

- **Machine-Speed Mitigation (Remediate):** Once an exploitable, high-priority risk is validated, autonomous code remediation layers—operationalized through autonomous security agents like CodeMender—step in to analyze the underlying architecture. These agents automatically generate, validate, and apply secure code patches directly to software libraries at machine speed. This effectively eliminates the human bottleneck,

[9] Defending Your Enterprise When AI Models Can Find Vulnerabilities Faster Than Ever | Google Cloud Blog
[10] Introducing Google AI Threat Defense | Google Cloud Blog

allowing critical systems to self-heal before an agile adversary can capitalize on a known vulnerability.

- **Continuous Detection (Monitor):** Even with a hardened foundation, true resilience across critical infrastructure requires constant vigilance during runtime. While pre-deployment, code-level scanning pipelines are excellent at catching flaws before software is pushed live, they are fundamentally incapable of blocking an active, zero-day exploit in real time. To counter this, our framework shifts defensive operations away from solely manual, human-paced oversight and toward machine-speed detection and real-time behavioral defense guided by frontline threat intelligence. Entities should utilize specialized, autonomous agents to triage suspicious behavior, and respond to live network intrusions at machine speed and an agentic Security Operations Center (SOC) functionality, to automate the detection, investigation, and tracking of emerging anomalies across their complex network, identity, and application telemetry.

By seamlessly tying real-world network exposure to automated, intelligent patch generation, we are establishing a comprehensive blueprint for modern cyber resilience. This continuous loop ensures that private and public entities can finally outpace the scaling volume and velocity of AI-driven adversarial operations.

Securing Artificial Intelligence through Bold and Responsible Innovation

We believe our approach to the frontier of artificial intelligence must be both bold and responsible. This means developing and deploying technology in a way that maximizes positive societal benefits while proactively engineering systems to withstand and mitigate modern adversarial pressures. For more than 20 years, Google has pioneered a Secure by Design approach, meaning we embed security into every phase of the software development lifecycle - not just the beginning or the end. Guided by Google's core AI Principles—originally published in 2018 and systematically updated to address the changing technology ecosystem—we design our AI systems from the ground up with robust security measures and strict safety guardrails.

Google's software and AI development pipeline relies on advanced threat modeling to proactively identify emerging threat trends and systemic risks, and to explicitly design our products for inherent safety. Rather than treating security as an afterthought, we continuously enhance safeguards inside our active products to offer scaled, adaptive protections to enterprise users and critical infrastructure operators across the globe.

GTIG partners closely with Google DeepMind to feed lessons learned from countering malicious activity directly into engineering processes. This creates positive feedback loops and

allows us to continuously improve the baseline safety and security of our AI foundation models. These intelligence-driven enhancements are applied dynamically at both the input/output classifier levels and deep within core model architecture. This rapid integration process is essential to maintaining extreme agility in our defensive postures and preventing sophisticated threat groups from abusing our technologies.

Conclusion

Cybersecurity has never been an environment where absolute perfection is possible. It will remain a fiercely contested, highly dynamic domain for years to come, demanding continuous innovation, speed, and structural agility to defeat adaptive adversaries.

As this Committee looks to secure our homeland and fortify the digital architecture supporting American critical infrastructure, Google stands ready to serve as a committed, transparent partner. By combining public-sector authority with private-sector technical innovation, we can harness the immense potential of artificial intelligence to tip the scales of cybersecurity permanently in favor of the defender.

Thank you for the opportunity to testify today. I look forward to answering your questions.

***

FRONTIER MODEL FORUM

Testimony of Chris Meserole
Executive Director, Frontier Model Forum

For the U.S. House of Representatives Committee on Homeland Security
Subcommittee on Cybersecurity and Infrastructure Protection

Hearing on "The AI Security Landscape: How Frontier Models, Agentic AI, and AI Coding Tools Are Reshaping Cybersecurity and Critical Infrastructure Resilience"

June 4, 2026

Introduction

Chairman Ogles, Ranking Member Ramirez, distinguished Members of the subcommittee, thank you for the opportunity today to testify on "The AI Security Landscape." With frontier AI capabilities advancing rapidly, understanding their impact on the resilience of U.S. cybersecurity and critical infrastructure has become increasingly important. I commend the subcommittee for its sustained attention to the issue and am honored to speak with you all this morning.

As Executive Director of the Frontier Model Forum, I am keenly aware of the security challenges and opportunities posed by frontier AI. The FMF is an industry-supported non-profit whose core mission is to bring together leading AI developers to advance frontier AI safety and security. Since our founding nearly three years ago, we have worked closely with our six member firms -- Amazon, Anthropic, Google, Meta, Microsoft, and OpenAI -- to execute on that mission and collectively address large-scale risks to public safety and security, including and especially risks from advanced cyber capabilities. To date, we have successfully established a first-of-its-kind information-sharing mechanism for frontier AI threats, allocated more than $10 million in funding for leading-edge AI security research and evaluations, and pioneered novel risk management practices and standards for frontier AI.

Notably, we undertook that work in anticipation of the AI capabilities we see today. I joined the FMF because I believed deeply that as frontier models and agents became more powerful, we would need trusted and credible channels by which information about frontier AI capabilities and risks could be shared out from industry to the broader public and policy community.[1] Now

[1] The FMF shares information about frontier capabilities, risks, and safety and security practices publicly through its website. It also maintains a private information-sharing channel for sharing information about frontier AI capabilities, vulnerabilities, and threat intelligence. See our recent update for more details on the latter: https://www.frontiermodelforum.org/updates/progress-update-fmf-information-sharing-of-frontier-ai-threats-and-vulnerabilities/.

that frontier models have demonstrated advanced cyber capabilities, I am even more convinced of the importance of maintaining such channels – which is why my comments today will refrain from advocating or lobbying for specific policy measures, and will instead aim to inform public debate and discussion of the security challenges and opportunities posed by frontier AI capabilities.[2]

My comments will center on three main points. The first is that the advanced cyber capabilities of today's frontier models follow a longstanding trendline and do not represent an unexpected jump in capability. The second is that the advanced cyber capabilities of today's models pose credible risks to cybersecurity and critical infrastructure, especially given the rise of adversarial distillation. Finally, the last point I'll make is that there is a great deal we can do to manage those risks, particularly when it comes to leveraging AI for cyberdefense, advancing cyber practices and standards, and building on existing information-sharing mechanisms and infrastructure.

Frontier AI Capabilities Remain on Trend

The most recent generation of frontier AI models have demonstrated impressive cyber capabilities. Although models like Anthropic's Mythos and OpenAI's GPT 5.5 are general-purpose models trained to carry out a wide range of tasks, they have exhibited extraordinary performance on cyber-related tasks specifically. Mythos, for example, became the first model to complete every task in CyBench, a widely-used cyber benchmark that measures an agent's ability to autonomously identify and exploit vulnerabilities.[3] Likewise, a version of GPT 5.5 recently solved a complex cyber range developed by the UK AI Security Institute and had the highest success rate per token across a range of autonomous tasks.[4] The evidence is clear: frontier agents can now autonomously perform complex cybersecurity tasks, including identifying novel vulnerabilities.[5]

Yet it is important to note that the impressive performance of the latest frontier models does not reflect a sudden jump in model capability. Instead, the improved capabilities of the most recent models are in line with long term trends. For example, in early 2025 a team of AI security researchers looked at the trendline on SWE-bench Verified, a challenging software engineering benchmark, and predicted scores would rise from roughly 60% at the beginning of that year to 87% in early 2026.[6] The forecast was largely on point: the leading models at the beginning of

[2] The FMF is registered as a 501(c)6 and has a policy against lobbying. For more on our governance structure and policies, see here: https://www.frontiermodelforum.org/about-us/#governance.
[3] See Anthropic, "System Card: Claude Mythos Preview" (April 7, 2026), pages 48-49. Accessed at https://cdn.sanity.io/files/4zrzovbb/website/7624816413e9b4d2e3ba620c5a5e091b98b190a5.pdf.
[4] UK AI Security Institute, "Our evaluation of OpenAI's GPT-5.5 cyber capabilities" (April 30, 2026). Accessed at https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities.
[5] Brian Grinstead and Christian Holler, *Mozilla*, "Hardening Firefox with Anthropic's Red Team" (March 6, 2026). Accessed at https://blog.mozilla.org/en/firefox/hardening-firefox-anthropic-red-team/.
[6] Govind Pimpale, Axel Højmark, Jérémy Scheurer, and Marius Hobbhahn, *arXiv*, "Forecasting Frontier Language Model Agent Capabilities" (March 3, 2026). Accessed at https://arxiv.org/abs/2502.15850.

this year – Gemini 3.1, Opus 4.6, and GPT 5.4 – all scored around 80%. Anthropic's reported score for Mythos of 93% is only slightly above where capabilities were projected to be by this time.

Likewise, the capabilities of today's most advanced models to find and detect real-world vulnerabilities on their own are also on trend. In the summer of 2024, a DARPA challenge used an LLM-based agent to autonomously find and patch a vulnerability in the common open-source database SQLite 3.[7] In the late spring of 2025, a team of UC Berkeley researchers using the leading frontier models of the time – GPT 4.1 and Sonnet 3.7 – found 15 unique zero-day vulnerabilities in existing open-source repositories.[8] Later that summer, Google revealed their Big Sleep agent had autonomously discovered roughly 20 zero-day vulnerabilities in widely-used open-source repositories.[9] The ability of the latest frontier agents to find high-severity vulnerabilities at scale is in keeping with an increasing trendline that has been underway for nearly two years.

The reason I'm underscoring the overall trend is not to diminish the importance of the latest frontier capabilities. Rather, it's to highlight that those capabilities should not have come as a surprise. To the extent that policymakers and the public were caught off guard by the most recent models, it should serve as a wake-up call: as I note below, we should use the current moment to strengthen existing public-private partnerships and information-sharing channels. With the same trendlines set to continue, policymakers will need to remain alert to the challenges they pose and the opportunities they present.

Frontier Risks to Cybersecurity and Critical Infrastructure

The cyber capabilities of advanced general-purpose models are inherently dual-use and can benefit both attackers and defenders. Agents that can autonomously detect zero-day vulnerabilities in critical software benefit society when they are controlled by responsible cyberdefenders, but pose significant risks if used by malicious actors. The more capable agents are at both identifying and exploiting vulnerabilities, the greater the security threat they may pose to public safety and critical infrastructure.

And the threat is real. Although cyber risks are easy to sensationalize and overstate, we know that malicious actors have already started to leverage the cyber capabilities of frontier models.

---

[7] Hanqing Zhao, "Autonomously Uncovering and Fixing a Hidden Vulnerability in SQLite3 with an LLM-Based System" (August 28, 2024). Accessed at https://team-atlanta.github.io/blog/post-asc-sqlite/.

[8] Zhun Wang, Tianneng Shi, Jingxuan He, Matthew Cai, Jialin Zhang, and Dawn Song, arXiv, "CyberGym: Evaluating AI Agents' Cybersecurity Capabilities with Real-World Vulnerabilities at Scale" (June 3, 2025). Accessed at https://arxiv.org/pdf/2506.02548v1.

[9] Heather Adkins, Google's Vice President for Security, published a public note on August 4, 2025 that "we are proud to announce that we have reported the first 20 vulnerabilities discovered using our AI-based "Big Sleep" system powered by Gemini." Accessed at https://x.com/argvee/status/1952390039700431184.

Threat actors linked to China, Iran, Russia, and North Korea have used advanced agents to carry out operations across the cyber attack lifecycle, including reconnaissance, exploitation, lateral movement, and data exfiltration.[10] Last fall one threat actor even developed the first known case of malware that used "Just-In-Time" AI, calling general-purpose models to generate malicious functions on demand.[11] Yet advanced AI has also enabled low-skilled actors as well. For example, relatively unsophisticated cyber criminals have started to engage in "vibe hacking" profitably, including through the development and sale of AI-generated ransomware.[12]

Given the capabilities and intent of known threat actors, the advanced cyber capabilities of existing models have significant implications for the resilience of U.S. cybersecurity and critical infrastructure. This is particularly true for small and under-resourced operators within critical sectors like water, energy, healthcare, and local government. For targets with thin defenses and outdated protocols, the vulnerability discovery and exploitation capabilities of frontier AI are likely to be especially impactful. Such targets may not have been worth the effort for skilled attackers to exploit in the past, but that will likely shift as it becomes easier to automate more and more of the cyber attack lifecycle. As noted below, leveraging AI for cyberdefense will become increasingly essential as a result.[13].

Threats from Adversarial Distillation

Critically, all of these threats are compounded by adversarial distillation. Although there are many legitimate uses for distillation itself, when it is carried out at industrial scale and outside a

[10] See Google Threat Intelligence Group, "GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools" (November 5, 2025) Accessed at https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools. See also Anthropic, "Threat Intelligence Report: August 2025" (August 2025). Accessed at https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf.
[11] Google Threat Intelligence Group, "GTIG AI Threat Tracker: Advances in Threat Actor Usage of AI Tools" (November 5, 2025). Accessed at https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools.
[12] Anthropic, "Threat Intelligence Report: August 2025" (August 2025), pages 4 and 15. Accessed at https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf.
[13] Which communities of attackers will benefit the most from advanced AI capabilities remains an open question. See Andrew Lohn, Center for Security and Emerging Technology, "Anticipating AI's Impact on the Cyber Offense-Defense Balance" (May 2025). Accessed at https://cset.georgetown.edu/publication/anticipating-ais-impact-on-the-cyber-offense-defense-balance/.

model provider's terms of service, distillation introduces significant safety and security risks.[14] Training a "student" model on the outputs of a more powerful "teacher" model enables the capabilities of the latter to be transferred, but not the associated safeguards and security mitigations.

The threat to cybersecurity and critical infrastructure is twofold. The first is obvious: if left unchecked, foreign rivals can leverage adversarial distillation to accelerate their own domestic AI capabilities, which state-linked actors can then use to target the United States.[15] The second is less straightforward: when adversarially distilled models are openly released, malicious actors of all kinds are able to leverage their capabilities for misuse without worrying about safeguards disrupting their efforts.[16] Any effort to secure U.S. critical infrastructure will be ineffective without a parallel effort to address adversarial distillation.

What Can Be Done

While we should remain sober and clear-eyed about the threats posed by advanced cyber capabilities, we should be equally clear-eyed about the many tools at our disposal to manage them. From leveraging AI for cyberdefense and strengthening information-sharing channels to accelerating the development of cybersecurity evaluations and standards, we have a wide array of methods and opportunities to improve the resilience of U.S. cybersecurity and critical infrastructure.

AI for Cyberdefense

One of the most effective responses to AI-enabled cyber threats is to ensure that defenders can leverage advanced cyber capabilities first. Many of the advances that make frontier AI useful for offensive cyber operations are similarly valuable for strengthening cybersecurity, particularly with respect to the following:

---

[14] See Anthropic, "Detecting and Preventing Distillation Attacks" (February 23 2026). Accessed at https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks; Google Threat Intelligence Group, "GTIG AI Threat Tracker: Distillation, Experimentation, and (Continued) Integration of AI for Adversarial Use" (February 12, 2026). Accessed at https://cloud.google.com/blog/topics/threat-intelligence/distillation-experimentation-integration-ai-adversarial-use; OpenAI, "Letter to the House Select Committee on the Strategic Competition between the United States and the Chinese Communist Party on Updated Stakes for American-Led, Democratic AI" (February 12, 2026). Accessed at https://assets.bwbx.io/documents/users/iqjWHBFdfxIU/rRmqI_jJcxb4/v0.

[15] Jared Dunnmon, Avanika Narayan, and Jon Saad-Falcon, *Foreign Affairs*, "China's AI Heist: How to Counter Beijing's Unauthorized 'Distillation'" (May 29, 2026). Accessed at https://www.foreignaffairs.com/china/chinas-ai-heist.

[16] Deepseek, which has been accused of adversarially distilling US frontier models, has been used to carry out cyber attacks. See Matt Pearl, Julia Brock, and Anoosh Kumar, *CSIS*, "Delving into the Dangers of DeepSeek" (February 24, 2025). Accessed at https://www.csis.org/analysis/delving-dangers-deepseek.

- **Proactive Vulnerability Discovery and Remediation.** Frontier AI systems can act as "friendly attackers," identifying vulnerabilities in software and networks before malicious actors can exploit them. Beyond finding weaknesses, advanced agents can increasingly help prioritize remediation efforts, generate patches, test fixes, and accelerate deployment.[17] As vulnerability discovery becomes more automated, AI-assisted remediation will be essential to maintaining secure systems.

- **Enhanced Detection and Incident Response.** For cyberdefenders, parsing the large volume of cybersecurity alerts and related telemetry can be overwhelming. Frontier agents can help analyze large quantities of security information, identify suspicious patterns, prioritize the most significant threats, and rapidly generate detection rules for newly discovered vulnerabilities and attack techniques. These capabilities can substantially improve the speed and effectiveness of defensive operations, particularly for organizations with limited cybersecurity resources.

- **More Secure Software Development.** AI should be used throughout the software development cycle to improve security. AI tools can identify insecure coding practices, flag potential vulnerabilities during development, and help engineers write safer code from the outset. As others have noted, AI agents can also be used to modernize legacy systems by rewriting older software in more secure, memory-safe programming languages.

Leading AI developers have already begun deploying advanced cyber capabilities in support of defenders. Recent initiatives such as Anthropic's Project Glasswing[18] and OpenAI's Trusted Access for Cyber[19] program focus on enabling trusted cybersecurity practitioners to identify and remediate vulnerabilities more quickly, improving defensive tooling for incident response, and making advanced cyber capabilities available to public-interest and critical infrastructure security efforts.[20] These efforts demonstrate that the same capabilities that raise concerns about offensive misuse can also be harnessed to strengthen collective cyber resilience.

Information Sharing

As noted above, the cyber capabilities demonstrated by today's frontier AI systems are in keeping with a longstanding trendline. Yet policymaker awareness about the overall trajectory of advanced cyber capabilities, and how recent models compare to it, has been relatively low.

---

[17] Justin W. Lin et al, *arXiv*, "Comparing AI Agents to Cybersecurity Professionals in Real-World Penetration Testing" (March 3, 2026). Accessed at https://arxiv.org/pdf/2512.09882.
[18] Anthropic, "Project Glasswing: An Initial Update" (May 22, 2026). Accessed at https://www.anthropic.com/research/glasswing-initial-update.
[19] OpenAI, "Introducing Trusted Access for Cyber" (February 5, 2026). Accessed at https://openai.com/index/trusted-access-for-cyber/.
[20] Additional industry efforts include Microsoft's multi-model agentic scanning harness (MDASH), focused on vulnerability discovery.

To increase public knowledge of AI capabilities and enable timely coordination and response to emergent risks, the following priorities are in order:

- **Strengthen Existing Information-Sharing Channels**. Many frontier AI developers already have bilateral voluntary channels with the U.S. government, including the Center for AI Standards and Innovation.[21] Likewise, all members of the FMF are party to our voluntary information sharing side agreement, which was explicitly designed to facilitate sharing about vulnerabilities, threats and concerning capabilities that are unique to the most advanced models and agents.[22] Each of these channels compliment the critical work and information-sharing mechanisms of established ISACs, ISAOs, and Sector Coordinating Councils. Rather than creating entirely new channels, policymakers should focus on ensuring that existing channels can effectively incorporate information related to frontier models and agents. Building on trusted relationships will be faster and more effective than standing up new institutions from scratch.

- **Provide Clear Guidance for Industry Engagement**. Effective information sharing within industry depends on clear guidance about what types of security-related information can be shared under existing antitrust and export control law.[23] Similar clarity should be provided regarding the treatment of sensitive information provided voluntarily to government agencies, including the circumstances under which it may be subject to public disclosure.

- **Focus on Signal Rather than Volume**. For frontier AI risks, channels should prioritize information that is meaningful and actionable. A focused approach helps to ensure that the most important signals are not lost amid a growing volume of technical information.

The FMF's own information-sharing efforts were established with similar priorities in mind. Our goal is to help support effective, robust and mutually reinforcing information channels across the AI ecosystem that improve safety knowledge and enable timely response to emerging issues and incidents.

Cybersecurity Practices and Standards

[21] Sanya Mansoor, *Guardian*, "US and tech firms strike deal to review AI models for national security before public release" (May 5, 2026). Accessed at https://www.theguardian.com/technology/2026/may/05/commerce-department-ai-agreements-google-microsoft-xai.

[22] Frontier Model Forum, "Progress Update: FMF Information Sharing of Frontier AI Threats and Vulnerabilities" (February 16, 2026). Note that our information sharing mechanism covers information not only about cyber risks, but CBRN threats and related risks to public safety and security. Accessed at https://www.frontiermodelforum.org/updates/progress-update-fmf-information-sharing-of-frontier-ai-threats-and-vulnerabilities/.

[23] Greater clarity and guidance on adversarial distillation specifically would be especially valuable.

Securing frontier AI does not require starting from scratch. The most effective approaches build on established cybersecurity principles and frameworks, while adapting or updating them as needed for the capabilities of advanced models and autonomous agents. The following priorities can help guide that effort:

- **Build on Established Cybersecurity Practices.** As the FMF has noted, the novelty of frontier AI capabilities does not demand novel security practices. Many of the important security controls for frontier AI are foundational to all cybersecurity: defense-in-depth architectures, internal security reviews, penetration testing, red-teaming, access controls, and continuous monitoring all play important roles in securing advanced AI systems.[24] As AI capabilities continue to advance, organizations should continue to strengthen and adapt these practices to address emerging threats.

- **Implement Layered Security.** AI agents are becoming more autonomous and more capable, which raises the stakes for agent security. As the FMF has observed in a recent issue brief, agent security breaks down across key layers: the underlying model, the harness built around it, the execution environment in which the agent operates, and the external tools the agent can invoke.[25] Each layer raises its own security questions and demands its own implementations. The model and harness introduce probabilistic, hard-to-predict failures that traditional security approaches were never designed to handle, while the execution environment and tool access lend themselves to deterministic controls that contain the damage when something goes wrong.

- **Consider Agent Scope.** Agents with greater scope can be used for a wider range of tasks, but that scope also extends the potential for harm if a system is compromised. Limiting the actions an agent can take only to what it needs to achieve a given task can minimize that harm. A useful frame is the "lethal trifecta": the risk of an attacker reaching private data increases when an agent combines three capabilities (access to private data, exposure to untrusted content, and the ability to communicate externally). Because agent security remains an open research problem, one mitigation for higher-risk uses is to ensure no single agent holds all three capabilities at the same time. The tradeoff is usability. Since tighter scope limits an agent's capability, deployers will need to balance between agent usability and security.

- **Advance Standards and Risk Management Frameworks.** Existing risk management practices and cybersecurity standards provide an important foundation for managing AI-related security risks. From recent and forthcoming guidance from NIST on agentic

[24] Frontier Model Forum, "Issue Brief: Foundational Security Practices" (July 31, 2024). Accessed at https://www.frontiermodelforum.org/updates/issue-brief-foundational-security-practices/.
[25] Frontier Model Forum, "Issue Brief: Emerging Security Practices for AI Agents" (June 3, 2026). Accessed at https://www.frontiermodelforum.org/issue-briefs/emerging-security-practices-for-AI-agents/.

security[26] to ongoing efforts to standardize frontier AI frameworks,[27] continued work to refine and operationalize risk management practices and standards in light of frontier AI capabilities will be essential.

The FMF has made this work a priority, publishing technical reports and best-practice guidance on frontier AI security and contributing to standards efforts so that emerging norms keep pace with rapidly advancing capabilities.[28]

Cybersecurity Benchmarks and Evaluations

Effective measurement is critical to effective risk management practices and policies. By providing standardized and repeatable tasks, cybersecurity benchmarks allow researchers to compare models, track capability improvements over time, and identify emerging trends. Sustaining credible measurement will require continued investment on several fronts:

- **Ensuring Benchmarks Stay Ahead of the Frontier**. Many publicly available cyber benchmarks are approaching saturation, with many models now scoring so highly on existing benchmarks that the results are no longer able to provide meaningful signals about improvements in capability. Developing more challenging evaluations will be critical to ensuring that measurement keeps pace with capability advances.

- **Grounding Evaluations in Real-World Threat Models**. Although measuring abstract cyber capabilities is useful, what matters most is whether those capabilities translate into meaningful security risks. Future evaluations should therefore be designed around realistic threat models and operationally relevant tasks, including vulnerability discovery, exploitation workflows, persistence, and other activities associated with real-world cyber operations. Doing so will help ensure that evaluation results provide meaningful insight into risks to public safety and critical infrastructure.

- **Expanding the Evaluation Ecosystem**. No single benchmark or evaluation methodology can fully capture the cyber capabilities of frontier AI systems. Robust assessment will require a diverse ecosystem that includes automated benchmarks, expert-led exercises, agentic evaluations, red-teaming, and uplift studies designed to measure how models affect the performance of human operators that are appropriate to the relevant threat

[26] National Institute of Standards and Technology, "Announcing the "AI Agent Standards Initiative" for Interoperable and Secure Innovation" (February 17, 2026). Accessed at https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure.
[27] INCITS, "INCITS 594-202x: Information Technology - Framework for Managing Unique Risks from Frontier AI" (May 14, 2026). Accessed at https://standards.incits.org/higherlogic/ws/public/projects/4370/details.
[28] Frontier Model Forum, "Technical Report: Managing Advanced Cyber Risks in Frontier AI Frameworks" (February 13, 2026). Accessed at https://www.frontiermodelforum.org/technical-reports/managing-advanced-cyber-risks-in-frontier-ai-frameworks/.

model. A broader evaluation ecosystem will provide a more comprehensive understanding of both capability development and associated risks.

Thankfully the US CAISI and many others, including the FMF, have already made important investments in advancing cybersecurity testing and evaluation. Building on those efforts will help ensure that policymakers, researchers, and critical infrastructure operators have access to reliable and timely information about the cyber capabilities of the latest frontier models and agents.

Conclusion

The latest generation of frontier AI models and agents have demonstrated impressive cyber capabilities, with significant implications for US cybersecurity and critical infrastructure resilience. Yet those capabilities are not unexpected: they are the continuation of a trend that has been visible for several years and that is likely to continue in the years ahead.

Addressing the risks posed by those capabilities will be challenging, but fortunately we have a strong foundation to start from. By building on and strengthening existing information-sharing channels and extending and updating longstanding cyber practices and frameworks, as well as leveraging AI for cyberdefense and investing in novel cyber evaluations, we can improve the resilience of our cybersecurity and critical infrastructure.

At the FMF, we are committed to advancing that effort through our work on risk management practices and standards, our support for scientific research, and our investment in information-sharing. Thank you for the opportunity to contribute my perspective this morning. I welcome the committee's attention to the security risks posed by frontier AI and look forward to answering your questions.

Written Testimony of Jack Cable
CEO & Co-Founder
Corridor

Before the U.S. House Committee on Homeland Security
Subcommittee on Cybersecurity and Infrastructure Protection

Hearing on
“The AI Security Landscape: How Frontier Models, Agentic AI, and AI Coding Tools Are Reshaping Cybersecurity and Critical Infrastructure Resilience”

June 4, 2026

Chairman Garbarino, Ranking Member Thompson, Chairman Ogles, and Ranking Member Ramirez, thank you for the opportunity to testify today.

My name is Jack Cable. I am the CEO and Co-Founder of Corridor. Our mission is to prevent a new and widespread wave of security vulnerabilities by securing AI coding. Before this, I helped build the Secure by Design initiative at CISA and was a top-ranked ethical hacker.

We are living in a time of profound change in cybersecurity. Coding agents, not software engineers, are increasingly the ones writing code and they're becoming more autonomous every day. They're creating new code at rates we've never seen before and, without appropriate guardrails, will introduce more vulnerabilities than ever before. This monumental change in how software is written is happening at the same time as frontier models like Mythos are increasingly capable of finding and exploiting vulnerabilities.

But before I talk about what's changing, I'd like to talk about what's staying the same. For one, attackers generally aren't exploiting new kinds of vulnerabilities. They're exploiting the same old vulnerabilities we've known about for decades. Even Mythos, which is an incredibly strong model, is surfacing issues like buffer overflows, first discovered back in 1972.

Similarly, long-standing defense mechanisms are still relevant. The case we built during my time at CISA around Secure by Design is more relevant than ever. Taking steps like refactoring code to use memory-safe languages will yield dividends for years to come.

The central challenge is not that AI creates entirely new categories of vulnerabilities. It is that AI dramatically increases the speed and scale at which old vulnerabilities can be introduced, found, and exploited. That means our response must shift from patching individual bugs to preventing entire classes of vulnerabilities at the source.

Today, I'll make three key points:

1. Hackers have more powerful tools than ever before.
2. AI is the dominant code writer today, and with scale comes more vulnerabilities.
3. Properly guided, AI coding represents the possibility of a more secure future.

Getting Ahead of the Bugpocalypse

Frontier models are becoming increasingly capable of performing complex cybersecurity tasks. Mythos and GPT-5.5 are just the latest iteration of models that – starting in earnest last fall – could perform exploits using robust end-to-end attack chains.[1] I want to be clear that these

[1] https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities

models aren't just hype – they are truly starting to rival, or exceed, humans on security tasks, and can do so at an unprecedented scale.

We will not be able to patch our way out of these issues. Of the over 1,500 vulnerabilities Anthropic has disclosed via Mythos, only 6% have been fixed (as of May 22, 2026).[2] Instead, we must get ahead of vulnerabilities at the source, which requires shifting focus to vulnerability prevention and wide-scale remediation.

The open question is whether each model release, for instance what we've seen with Mythos or GPT-5.5, will discover incrementally more vulnerabilities, or if the number of vulnerabilities will start to plateau after some point.

The evidence so far is mixed. Mozilla released fixes for 271 vulnerabilities in the Firefox browser attributed to Mythos, after having discovered 22 vulnerabilities with Opus 4.6.[3] But Curl, a widely used open source library, reported discovering just one new vulnerability with Mythos.[4]

The challenge is especially acute for open source software, since it's a public good. Open source software underpins every software service we rely upon, including across critical infrastructure and the Federal government. A string of recent incidents have highlighted the need for sustained investment in open source software security. In this new era, the government has a crucial role to play in providing funding for the maintenance and securing of this software infrastructure.[5]

Regardless of whether subsequent model versions will continue to surface drastically more vulnerabilities, they still won't be – by and large – discovering novel vulnerability classes. Thus the best approach is to ensure that software is built secure by design – that code generated is resilient to known attack patterns and hard to exploit in the first place. This is where AI coding comes in.

Securing AI Coding

The software industry is being truly revolutionized by AI coding. The most productive engineers no longer write code – they instruct fleets of AI agents to do so on their behalf. These coding agents can write high volumes of code significantly faster than ever before and, increasingly, they can do so with less human oversight.

---

[2] https://red.anthropic.com/2026/cvd/
[3] https://blog.mozilla.org/en/privacy-security/ai-security-zero-day-vulnerabilities/
[4] https://daniel.haxx.se/blog/2026/05/11/mythos-finds-a-curl-vulnerability/
[5] https://www.cisa.gov/news-events/news/lessons-xz-utils-achieving-more-sustainable-open-source-ecosystem

The code hosting platform GitHub has reported a 14x increase in code committed in 2026 than in 2025.[6] Sundar Pichai has said that 75% of the new code at Google is AI generated.[7] For our own development at Corridor, coding agents write the vast majority of our code. Today, you can start a coding agent from your phone, and it'll work for an hour and produce a significant code change while you eat lunch. Coding agents aren't yet perfect, but they represent the biggest shift in software development in decades and are getting better every day.

This comes with profound implications for security. The traditional code review model depends on human code review and imprecise tooling. This model is breaking down under the sheer volume of code now being generated by coding agents. Companies are beginning to deploy code without human review, as that is now the main constraint to moving quickly.

While coding agents are better at writing secure code on a per-line basis than humans, they still often introduce vulnerabilities. These vulnerabilities scale drastically with the amount of code these agents produce. According to the academic benchmark BaxBench, around a third of code written by the best coding models has security or correctness issues.[8] Internal Corridor data suggests that 13% of code changes created by coding agents have security vulnerabilities.

We are stuck between a rock and a hard place. Mythos-level models on the offensive side are better at finding and exploiting vulnerabilities than any tool in the history of hacking. Coding agents are writing more code than ever before, introducing an increasing number of security vulnerabilities that human software engineers can't keep up with. To prevent the bugpocalypse, we need a new path forward.

Properly Guided, AI Coding Represents the Possibility of a More Secure Future

The good news is that, in our work at Corridor, we are seeing that AI coding agents can follow security instructions better than most humans. Across our customers, we see a 60% reduction in vulnerabilities by giving specific context to the coding agent at the planning stage. Similarly, academic studies have found that optimized security-specific context improves the ability of agents to write secure code.[9]

At Corridor, we work with companies at the cutting edge of AI-powered software development to prevent vulnerabilities at the time of code generation and analyze code before it's deployed in the real world. That means we directly intervene as AI coding agents do their work, guiding them based on years of best practices and discovered vulnerabilities and have AI do a thorough review

[6] https://x.com/kdaigle/status/2040164759836778878
[7] https://www.semafor.com/article/04/24/2026/google-ceo-says-75-of-companys-new-code-is-ai-generated
[8] https://baxbench.com/
[9] See https://arxiv.org/pdf/2605.08382 and https://baxbench.com/

of code before it's pushed to production systems. This is based on a continually-updated model of a company's security posture.

For existing code, frontier models can accelerate security-oriented refactors. What used to cost millions of dollars and years of human effort can now be accomplished for thousands of dollars in weeks. Initiatives like DARPA's TRACTOR program, focused on translating unsafe code to use memory-safe languages, are crucial.[10] Rather than playing whack-a-mole every time a new model comes out, companies need to invest now in the foundational security practices that will make their code more resilient against any type of adversary.

Open-weight models

In addition to the frontier closed-weight models we've discussed, open-weight models are increasingly capable at coding, security, and other tasks. Models like Kimi and Qwen rival the performance of frontier models released 3-6 months ago[11] at a fraction of the cost.

It's an inevitability that open-weight models will proliferate. We've already seen instances of model distillation, where creators of open-weight models train on output of closed-weight models.[12] It's generally futile to attempt to restrict model capabilities. The best defense against adversaries leveraging open-weight models to conduct attacks is to shore up foundational cyber defenses, such as building software to be secure by design and resilient to entire classes of vulnerabilities.

Comparatively, the core benefit of open-weight models is that you can use them as you wish. While with closed-weight models you are dependent on the model provider to host them, open-weight models allow you to host them yourselves. In addition to lower costs, open-weight models can be fine-tuned, which can be used to rival, and in some cases exceed, the performance of closed-weight models.

There are currently no frontier open-weight models from the United States. Both Meta and OpenAI have released open-weight models in the past, but these have quickly become outdated. The United States should build capacity for domestic models, and there are several promising non-profit initiatives developing open-weight models like Marin[13] and Olmo from the Allen Institute for AI.[14]

---

[10] https://www.darpa.mil/research/programs/translating-all-c-to-rust
[11] https://swe-rebench.com/
[12] https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks
[13] https://marin.community/blog/2025/05/19/announcement/
[14] https://allenai.org/olmo2

A thriving AI industry demands both the cutting-edge performance of closed-weight models, and the low-cost and flexibility of open-weight models. Without both, we will be less secure and fall behind as a nation.

Recommendations

To recap, we must stop creating new vulnerabilities, shore up the shared foundation adversaries will hit hardest, and maintain America's lead through open-weight models. My recommendations to the Committee are:

1. **Prevent vulnerabilities in new code:** The single highest-leverage step is to stop entire classes of vulnerabilities at the point of code generation rather than discovering them later. The government should start with its own house: Congress should enable AI coding among the Federal government and its contractors and require them to have security guardrails in place that prevent entire classes of vulnerabilities from being introduced.
2. **Harden the open source foundation.** Open source software underpins critical infrastructure and the Federal government alike, and as a public good it will be hit hardest by adversaries wielding frontier models. Rather than one-off fixes, Congress should establish and fund a multi-billion-dollar non-profit initiative focused on large-scale, security-oriented refactors[15] and maintenance of critical open source components – including, when needed, encouraging or sponsoring a fork[16] and helping to identify new maintainers of key open source projects.[17] I also encourage the Committee to bolster CISA's capabilities to partner with the open source ecosystem by passing The Securing Open Source Software Act, which would establish foundational expertise in government.
3. **Foster an Ecosystem of American-Made Open-Weight Models:** Lastly, we must advance frontier open-weight models from the United States. Recent moves like the NSF partnership with NVIDIA to support the Open Multimodal AI Infrastructure to Accelerate Science (OMAI) is a step in the right direction.[18] The U.S. Government should go even further and help support the development of open-weight models that can complement the cutting-edge closed-weight models from industry.[19]

Thank you for the opportunity to testify today. I look forward to your questions.

[15] https://www.linkedin.com/pulse/open-source-runs-world-shouldnt-run-goodwill-alone-jen-easterly-9loxe/
[16] A fork is a copy of an open source project.
[17] https://www.chainguard.dev/unchained/the-hardest-fork
[18] https://www.nsf.gov/news/nsf-nvidia-partnership-enables-ai2-develop-fully-open-ai
[19] https://foreignpolicy.com/2023/06/12/ai-regulation-technology-us-china-eu-governance/

Subcommittee on Cybersecurity and Infrastructure Protection

Committee on Homeland Security

Hearing:

"The AI Security Landscape: How Frontier Models, Agentic AI, and AI Coding Tools Are Reshaping Critical Infrastructure Security and Resilience"

Statement of Matthew Guariglia, Ph.D

Senior Policy Analyst

Electronic Frontier Foundation

June 4, 2026

As a Senior Policy Analyst for the Electronic Frontier Foundation, I thank the Chairs, Ranking Members, and Members of the Committee for the opportunity to share EFF's views on the benefits and harms associated with AI, particularly Generative AI, and options for responsible government interventions in its continued development and deployment. EFF is a nonprofit organization dedicated to protecting privacy, innovation, and free expression in the digital world. EFF is primarily funded by contributions from more than 30,000 dues-paying members. More than 80% of that funding consists of donations under $10,000. We receive less than five percent of our funding from corporate sponsors.

For 35 years, EFF has represented the interests of technology users, both in court and in policy debates, to help ensure that law and technology support and enhance our civil liberties.

AI is an all-encompassing term that seems to grow to include more technologies and use cases by the day. It is vital that we are clear about what we mean when discussing these tools in a cybersecurity context. For example, the use of general-purpose AI models to compile all possible information on citizens from a variety of sources poses a significant threat to privacy. By contrast, the use of a niche model for a specific purpose, such as improving accessibility on a website for the vision-impaired or models tasked specifically with finding vulnerabilities in critical infrastructure, poses less risk to privacy and civil liberties.

AI tools can be useful for all manner of hobbyists, scholars, and businesses, but they can also be misused and misunderstood. And when they become part of critical national security and cybersecurity infrastructure, those risks only increase.

Accordingly, governments must not adopt emerging and powerful technologies without also adopting strong and clear safeguards to protect Constitutional rights. This is of particular urgency because of the demands recently put on Anthropic by the Pentagon to make their technology available for use for all purposes, including those it was not

designed for, like mass surveillance of Americans.[1] EFF opposes the use of generative AI for the purposes of mass government surveillance because that use supercharges unconstitutional violations of civil liberties and because government secrecy prevents the public and lawmakers from knowing when generative AI models make mistakes. That is the baseline upon which all the rest of our recommendations rest.

I. National Security

There is great temptation, motivated by both national interest and lobbying from for-profit enterprises, to deploy emerging technology as quickly as possible. In fact, the Pentagon is already making rapid strides to deploy AI models without the rigorous testing and trial periods done by previous administrations.[2] While the desire to gain a strategic and computational edge is understandable, this rapid deployment raises a number of concerns as to whether the Intelligence Community and the military are meeting their transparency, accountability, and civil liberties obligations.[3] What is worse, this rapid deployment may *also* undermine our national security infrastructure.

A. Supercharging Surveillance

In 2024, the Biden Administration issued a memorandum[4] declaring the intentions of the national security apparatus to leverage the private sector's AI expertise for the public benefit. Unfortunately, meeting that goal may involve merging proprietary and

---

[1] Matthew Guariglia, "*The Department of Defense Wants Less Proof its Software Works*," Electronic Frontier Foundation (October 31, 2025), available at https://www.eff.org/deeplinks/2025/10/department-defense-wants-less-proof-its-software-works

[2] Id.

[3] Matthew Guariglia, "*The U.S. National Security State is Here to Make AI Even Less Transparent and Accountable,*" Electronic Frontier Foundation (November 19, 2024), available at https://www.eff.org/deeplinks/2024/11/us-national-security-state-here-make-ai-even-less-transparent-and-accountable

[4] *"National Security Memorandum on Advancing the United States' Leadership in Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security and Trustworthiness of Artificial Intelligence"* White House (October 24, 2024), available at https://www.presidency.ucsb.edu/documents/national-security-memorandum-advancing-the-united-states-leadership-artificial

government secrecy. This level of opacity creates a system where training data, algorithmic decision-making processes, and use cases related to surveillance and data analytics, including against Americans, are inaccessible to plaintiffs, independent auditors, and anyone whose liberties are impinged by use of AI tools. It also means a system whereby illegal and unethical actions or mistakes and hallucinations enacted by AI might go unknown to the public.

For decades, EFF has fought[5] to expose and challenge secret government interpretations of national security statutes, under which the government has unconstitutionally spied on Americans and retained and analyzed that information. Between open-source intelligence methods like social media surveillance, purchasing intimate data like geolocation from the data broker market, and more traditional methods of signals intelligence which leverages data from telecommunications companies and internet providers, the Intelligence Community is able to access extraordinary amounts of data—not just from surveillance targets overseas but also from Americans who have Constitutional protections from warrantless and indiscriminate surveillance.

Artificial intelligence poses two big, interrelated problems in relation to decades of mass surveillance infrastructure and mostly classified legal interpretations. The first is the drastically increased capacity of AI to analyze information on behalf of the government. Modern AI can synthesize massive amounts of personal data into a comprehensive, exceedingly intimate portrait of an individual's private life, including their political affiliations, religious beliefs, personal communications, medical conditions, and sexual activities. For example, an LLM could infer an individual's association with a particular mosque based upon frequent visits to the mosque's website, engagement with the mosque's social media posts, and their cell phone's physical proximity to the mosque during religious services.

---

[5] Andrew Crocker and Aaron Mackey, "*Victory! EFF Wins National Security Letter Transparency Lawsuit,*" Electronic Frontier Foundation (May 14, 2019), available at https://www.eff.org/deeplinks/2019/05/victory-eff-wins-national-security-letter-transparency-lawsuit?language=el

In a strategic context, one can see many use cases where this would be a beneficial tool. From a civil liberties perspective, however, it becomes more problematic. AI analytic tools can easily combine information that can reveal sensitive personal information about an individual—without requiring preexisting probable cause. Before there was a smart phone in every pocket, many of our security protections relied on the high cost of surveillance. It took quite a lot of individual work to track any one person, and there was no possibility of historical data collection. With an assist from data brokers and government databases, AI has the potential to exponentially increase government capacity to track huge swaths of the population, exposing Americans to granular levels of surveillance with the click of a button.

These risks are already realities in other countries. For example, in China, the government uses AI to analyze and integrate vast amounts of personal information collected through social media monitoring, surveillance cameras, facial recognition systems, and other forms of surveillance. AI can use this information to identify dissent, allowing the Chinese government to locate dissidents and censor government criticism.[6] In a domestic policing context, real-time crime centers and the AI-enhanced platforms that fuse information from sometimes hundreds of sources are also making these fears a reality.[7] Use by the federal government and the data streams that accompany national security signals intelligence, including private digital communications harvested under FISA Section 702, would drastically exaggerate the problem.

The second major concern is the inadequate systems of human oversight. Time constraints, protocols, or disinterest prevent verification of determinations created by AI systems, exacerbate the inherent difficulties in ensuring accountability with a technology that can process large amounts of information. In many contexts, we are already experiencing the ramifications of mistakes in AI analysis and decision-making

---

[6] Darrell M. West, "*How AI Can Enable Public Surveillance*," Brookings Institute (April 15, 2025), available at https://www.brookings.edu/articles/how-ai-can-enable-public-surveillance/.

[7] Andrew Guthrie Ferguson, "*Real-Time Crime Centers and The Brady Puzzle,*" Boston University Law Review (forthcoming 2026), available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6468120

processes. Hallucinations, misreadings of inputs or training data, or overstepping guardrails are all common occurrences that are having a range of consequences for people in the United States. While fictional court cases generated by AI[8] may feel like a funny headline, other inaccuracies have had far less benign outcomes. From housing and employment to policing and immigration decisions, AI has already been integrated into many bureaucratic decision-making processes using often unknown or invasive data streams to make impactful choices for a person's life.[9]

"Rubber-stamping" AI decisions or conclusions, combined with inadequate verification procedures, creates a climate where often unaccountable and private AI systems are directing consequential government decisions rather than serving the will and priorities of civil servants. In the national security context, these types of mistakes and the lack of transparency and accountability that follows could have devastating effects on all of our safety.

B. Addressing Cybersecurity Risks

1) Understanding Frontier Models, Agentic AI, and AI Coding Tools

When most people think about AI, they are probably thinking about "generative" AI. Generative AI models—like ChatGPT, Google Gemini, or Claude—are computer systems that can respond to requests, or "prompts," in plain language. The responses can include a plain language response, or images, video, or audio. Or, importantly for the purposes of this Committee, the response can include the source code of a computer program.

---

[8] Pamela Langham, "*Massachusetts Lawyer Sanctioned for AI-Generated Fictitious Case Citations,*" Maryland State Bar Association (March 4, 2024), available at https://www.msba.org/site/site/content/News-and-Publications/News/General-News/Massachusetts_Lawyer-Sanctioned_for_AI_Generated-Fictitious_Cases.aspx

[9] Adam Schwartz and Catalina Sanchez, "*Americans are Uncomfortable with Automated Decision-Making,*" Electronic Frontier Foundation (September 3, 2024), available at https://www.eff.org/deeplinks/2024/08/americans-are-uncomfortable-automated-decision-making

The types of generative AI models that can analyze and produce text or computer code are called large language models, or LLMs. Current LLMs are capable of analyzing complicated questions and producing meaningful responses. The ones that can analyze the most complicated questions and produce the best answers are considered "frontier AI." As the AI industry invests in new models, the "frontier" moves, so an AI model that is considered to be at the frontier today will probably not be at the frontier in a couple of months.

At the time of this hearing, five US companies are generally considered to be operating at the frontier of AI development. There is strong competition from the entire world to develop similar or better AI models. And even those models not on the frontier are remarkably capable.

Some tools go beyond answering questions and are also capable of performing actions related to those questions. This is called "agentic AI," and is a combination of text-generating LLMs with additional software that understands how to perform certain actions. Think of the LLM as a brain in a jar, and an agentic system as a faithful lab assistant that knows how to carry out certain commands on behalf of the brain.

For instance, an LLM by itself can answer the question "read the source code for this program and tell me about all the bugs." An agentic AI system that is configured with an action to send email can both answer questions and take action. For instance, an agentic AI system configured to send emails could take action in response to the prompt "find all the bugs in this source code, and email them to example@example.com."

Agentic AI is particularly effective in the context of AI coding tools. Plain generative AI can write a program, but might make mistakes. An agentic AI coding tool can write a program and also send a command to the operating system to run a compiler on that program. The compiler checks for programming errors. If it produces an error, the agentic AI coding tool can make changes to the program it wrote and try again until the compiler succeeds. If the compiler succeeds, the tool is done and can report success.

Agentic techniques allow AI coding tools to perform more complex tasks with a higher chance of success, without the need for much human interaction. In particular, finding software security vulnerabilities has historically been a high-skill task requiring insight and understanding of the code being studied. Recent advances in both the core models and the agentic "harnesses" they are used with have dramatically reduced the skill required to find vulnerabilities.

2. AI, Cybersecurity, and the Problem of Vulnerability Hoarding

As AI tools have gotten better at writing code, they have also gotten better at finding cybersecurity vulnerabilities in existing code. A vulnerability is a bug that allows an adversary to break a computer system in some way, creating, in turn, an opportunity to crash it, extract private data, or secretly take control of it.

Security experts—both attackers and defenders—search for software vulnerabilities. Attackers do so to exploit them. Defenders do so to fix them so they cannot be exploited in the future. In intelligence and security operations, government acts, at different times, as an attacker and a defender. In espionage, government sometimes exploits vulnerabilities to gain access and extract information. At the same time, government acts to protect its own citizens and critical infrastructure from attackers.

This dual role creates a tension: when a government agency discovers a vulnerability, should the agency hoard it for future exploitation or disclose so it can be fixed? The ubiquity and power of AI vulnerability research decisively tilts the scales in favor of disclosure. For example, the NSA found and developed a number of specific exploits, such as EternalBlue and EpMe/Jian,[10] and kept them open for surveillance purposes, only to have bad actors and adversarial nation states find and use those exploits.[11]

---

[10] "*Eternal Blue*," Wikipedia (last accessed June 1, 2026), available at https://en.wikipedia.org/wiki/EternalBlue

[11] Andy Greenberg, "*China Hijacked an NSA Hacking Tool in 2014 – and Used It for Years,*" Wired (February 22, 2021), available at https://www.wired.com/story/china-nsa-hacking-tool-epme-hijack/

This will happen again. A vulnerability is an observable fact and it is common for the same vulnerability to be independently discovered multiple times. As AI tools make vulnerability research cheaper, independent discovery will become more common and holding vulnerabilities secret will have less value. And at the same time, fixing them promptly will have much more value.

3. How the Government Can Help.

These are areas where government investment could support cybersecurity without intruding on individual rights. First, the government should fully commit to disclosing and helping to fix software bugs, rather than hoarding them for surveillance.

In addition, the government could help address “the patch gap.” When software publishers fix vulnerabilities, users of that software must still update to the latest version in order to receive the benefit. The time between a fix becoming available and a given user installing it is called the patch gap. A brief patch gap—updates installed promptly and reliably—is good for cybersecurity. A long patch gap is disastrous because it gives attackers plenty of time to exploit unpatched systems.

On far too many of our critical and governmental systems, the patch gap is long. Some systems may be entirely unpatchable. But for many systems, the government could support AI-powered security research by making fundamental investments in cybersecurity hygiene: updating all systems promptly, keeping accurate inventories, retiring unmaintained systems and software, and reducing the attack surface of all systems.

The government can also help reduce the national attack surface by investing in the development and deployment of memory safe software in critical systems. Different software development methodologies produce vulnerabilities at different rates. Choice of programming language in particular plays a huge role. Many modern programming languages invest in a property called “memory safety” that automatically detects and prevents certain common bugs. At least 65% of software vulnerabilities are due to the

lack of memory safety[12]. Longstanding programming languages C and C++, notably, do not have memory safety and have no credible path to adding it. In a world where vulnerabilities are easy to find, one of the best defenses is to have fewer vulnerabilities.

II. Other Remedies and Danger Mitigation Options

When the tech companies themselves are trying to insist on guardrails that the military is trying to override, it is up to Congress to step in and provide necessary and balanced regulations, for both the protection and continued security of the American people.[13]

A. Transparency

As noted, the government should share information about newly discovered vulnerabilities, threat models, and mitigation measures with Congress and the public.

Excessive secrecy—often attributable to overclassification and/or overbroad protections for vendors' commercial information—thwarts Congressional oversight and conceals risks to security and civil liberties.[14] Congress should not allow any administration to withhold information it needs to exercise its constitutional prerogatives. Members of Congress should have more access to information about what the Administration has done and is doing.

More broadly, Congress should create a statutory framework for classified information that would restore checks and balances to what has become a system of executive classification. Without greater transparency, Congress will lack the information it needs to identify regulatory gaps and pass carefully targeted laws that address true AI risks

---

[12] Alex Gaynor, *"What Science Can Tell Us About C and C++ Security,"* AlexGaynor.Net (May 27, 2020), available at https://alexgaynor.net/2020/may/27/science-on-memory-unsafety-and-security/

[13] *Supra Note 1*

[14] Faiza Patel and Patrick C. Toomey, *"Bring Transparency to National Security Uses of Artificial Intelligence,"* Just Security (April 4, 2024), available at https://www.justsecurity.org/94113/bringing-transparency-to-national-security-uses-of-artificial-intelligence/

without threatening technological development or the freedom to use tools for beneficial purposes.[15]

Greater transparency would also allow security researchers and other experts to analyze the government's cybersecurity protocols and identify aspects in need of improvement—before our adversaries do. Moreover, U.S. entities, such as defense contractors and other large U.S. companies, may be at risk of similar cyberthreats. Public disclosure of new vulnerabilities and emerging threats would allow the private sector to harden defenses and patch vulnerabilities before attacks occur.[16]

Finally, greater transparency enables better public accountability for potential cybersecurity failures, which can compromise individuals' personal information, jeopardize national security, and otherwise profoundly impact the public interest[17]. Given public distrust of both AI and government surveillance, transparency is essential to preserving trust and legitimacy in the eyes of constituents.[18]

B. Limit AI-Enabled Surveillance of Americans

Minimizing the collection of Americans' personal information and improving procedural safeguards against illegal surveillance are also imperative. That means placing sharp,

---

[15] Tori Noble, Katharine Trendacosta, and Kit Walsh, *"Smart AI Policy Means Examining Its Real Harms and Benefits,"* Electronic Frontier Foundation (February 4, 2026), available at https://www.eff.org/deeplinks/2026/02/smart-ai-policy-means-understanding-its-real-harms-and-benefits

[16] Katharine Megas, Angela Smith et al, *"The Importance of Transparency – Fueling Trust and Security Through Communication,"* Cybersecurity Insights, a NIST Blog (April 3, 2023), available at https://www.nist.gov/blogs/cybersecurity-insights/importance-transparency-fueling-trust-and-security-through ; see also Suzanne Spaulding, *"The Importance of Transparency in Cybersecurity: 2023 Security Summit at the Fortinet Championship,"* Fortinet (September 24, 2023), available at https://www.youtube.com/watch?v=wtCsG1LpRzA

[17] Marisol Cruz Cain et al,*"OMB Action Needed to Address Privacy-Related Gaps in Federal Guidance,"* United States Government Accountability Office (March 2026), available at https://files.gao.gov/reports/GAO-26-107681/index.html?_gl=1*163n46x*_ga*MTc1NDU0OTExNy4xNzgwMTA3Mjc2*_ga_V393SNS3SR*czE3ODAxMDcyNzYkbzEkZzEkdDE3ODAxMDg2MzckajU5JGwwJGgw

[18] Valerie Wirtschafter, *"For AI to make government work better, reduce risk and increase transparency,"* Brookings Institute (January 16, 2025), available at https://www.brookings.edu/articles/for-ai-to-make-government-work-better-reduce-risk-and-increase-transparency/

enforceable limits on government spying, requiring meaningful judicial oversight of surveillance, and passing laws that prohibit the government from bypassing the Fourth Amendment by buying personal information in bulk from data brokers and other private entities.

EFF urges Congress to take two immediate steps: First, enact the Fourth Amendment is Not For Sale Act, which passed the House of Representatives in 2024 by a vote of 219 – 199.[19] This bipartisan bill would prohibit the government from purchasing digital data on individuals it would otherwise need a warrant to collect. This is an especially important body of information that should be minimized in order to prevent it from invasive analysis by AI models in light of the Office of the Director of National Intelligence's creation of a streamlined marketplace for the Intelligence Community to purchase personal information from the data broker industry.

The second important step that Congress could take right now to curb the potential for negative impact of AI use on civil liberties is meaningful reform of mass surveillance authority Section 702 of the Foreign Intelligence Surveillance Act (FISA). Section 702 remains one of the major ways that the National Security Agency compels private companies to hand over the digital communications of an unknown number of U.S. persons—at least several thousand of which have been queried by the Federal Bureau of Investigation without a warrant.

C. Avoid Centralized Government Control Over AI Developers and Models

It is essential to preserve the rights of companies to conduct frontier AI research, freely develop multi-use models for the public, and publicly release new models without the need to ask the government for permission to do so. Extraordinary government interventions, like limits on models that companies may release, are neither appropriate nor necessary to address AI risks. Though AI may pose novel security risks, those risks can be effectively mitigated by hardening insecure government information systems.

---

[19] *Fourth Amendment Is Not For Sale Act, H.R. 4639,"* Congress.Gov (last accessed June 1, 2026), available at https://www.congress.gov/bill/118th-congress/house-bill/4639