FRONTIER MODEL FORUM

Written 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.