JOINT INFORMATIONAL HEARING

Assembly Select Committee on Cybersecurity &

Assembly Committee on Privacy and Consumer Protection

Frontier Artificial Intelligence and the Future of Cybersecurity

Monday, August 10th, 2026 — 11:00 a.m.

Capitol Room 447

BACKGROUND INFORMATION

OVERVIEW

Artificial intelligence (AI) is driving transformation across virtually every sector of society. One area undergoing changes is the cybersecurity landscape, where AI is accelerating both sides of the equation. While AI enables defenders to identify previously unknown vulnerabilities and streamline remediation efforts, it also allows malicious actors to discover and exploit weaknesses at unprecedented speed, transforming cybersecurity into a race between vulnerability discovery and defense.

In April, these emerging capabilities garnered widespread public attention when Anthropic's Claude Mythos Preview autonomously identified and exploited previously undiscovered vulnerabilities, including flaws that had remained undetected for decades, across major operating systems, web browsers, and other widely used software.[1] These concerns intensified last month during an internal OpenAI safety evaluation. Tasked with completing a test, the agent escaped its testing environment, accessed the open internet, and compromised Hugging Face, a widely used platform for hosting and sharing AI models and datasets in an effort to obtain answers to the benchmark.[2] As frontier AI becomes increasingly advanced, policymakers face the challenge of understanding how these technologies should be evaluated, governed, and responsibly deployed to strengthen California's cybersecurity resilience.

This informational hearing will:

- Examine how frontier AI is redefining offensive and defensive cybersecurity capabilities.
- Explore testing and benchmarking frameworks used by researchers to quantify model risks.
- Delve into California’s readiness to safeguard state infrastructure through public-private collaboration and proactive oversight.

---

[1] Anthropic, *Project Glasswing: Securing Critical Software for the AI Era* (Apr. 7, 2026).
[2] OpenAI, *OpenAI and Hugging Face partner to address security incident during model evaluation* (Jul. 21, 2026).

UNDERSTANDING ARTIFICIAL INTELLIGDENCE

Frontier artificial intelligence refers to the most advanced generation of foundation models currently available.[3] Foundation models are large-scale AI systems trained on vast quantities of text, source code, images, and other data. Rather than being designed to perform a single predefined task, these models can be adapted to perform a wide range of functions, including generating and analyzing text, writing and reviewing software code, and reasoning through complex technical problems.[4] These foundation models underpin many of the generative AI applications widely used by the public today, such as ChatGPT and Gemini. Today's leading frontier AI systems include models developed by Anthropic, OpenAI, Google DeepMind, and xAI.

Recent innovation has also led to agentic AI systems. Rather than simply responding to individual prompts, agentic AI systems have the capability to operate more like digital assistants. Users assign the AI system to a broader objective, and the system can then develop a step-by-step plan, gather and analyze information, use external software tools, write and execute computer code, and adapt its approach as new information becomes available. These capabilities represent a significant evolution from earlier generative AI systems, enabling frontier AI to complete increasingly complex workflows with limited human direction and oversight.[5]

These increasingly autonomous capabilities are also being demonstrated by experimental agentic systems developed by leading AI laboratories. OpenAI has begun evaluating advanced agentic models, including GPT-5.6 Sol and other internal research systems, to assess how increasingly autonomous AI agents perform complex real-world tasks. Anthropic's Claude Mythos Preview is a complementary research effort focused specifically on evaluating frontier AI's cybersecurity capabilities. The following section examines how agentic systems developed by OpenAI and Anthropic are reshaping the cybersecurity landscape and why they have attracted significant attention from researchers and policymakers.

FRONTIER AI IN CYBERSECURITY

Historically, cybersecurity has relied primarily on highly trained professionals to identify and defend against software vulnerabilities. Security researchers traditionally reviewed software code, tested computer systems for weaknesses, and investigated cyber incidents to uncover and remediate security flaws.[6] Today, frontier AI can assist with many of these same tasks by reviewing vast amounts of software code, identifying security vulnerabilities, helping developers write more secure software, generating code to test system defenses, and supporting aspects of cyber investigations in seconds.

---

[3] Dana Larson, *Understanding Frontier AI: Definitions, Models, and Business Impact* (May 26, 2026)
[4] Roya Firoozi et al., *Foundation Models in Robotics: Applications, Challenges, and the Future* (Dec. 13, 2023).
[5] Cole Stryker, *What Is Agentic AI?* IBM Think.
[6] SentinelOne, *History of Cybersecurity: Everything You Need to Know* (updated Jan. 26, 2026).

Another recent incident involving OpenAI demonstrated the power of agentic systems: during a cybersecurity evaluation in late July, OpenAI tested two advanced models with their ordinary safety restrictions reduced or disabled. Rather than limiting themselves to the vulnerabilities the benchmark was intended to test, the models discovered a previously unknown flaw in the proxy connecting their isolated testing environment to the internet, escaped that environment, and inferred that answers to the benchmark might be stored on Hugging Face. The models then independently chained together vulnerabilities, obtained credentials, commandeered outside computing resources, and penetrated portions of Hugging Face's production infrastructure in an effort to retrieve those answers.[7]

> 'The intrusion continued for several days, involved thousands of individual actions, and extended to accounts on several additional online services before Hugging Face detected it. A recent Politico article describes the scope of the incident: Hugging Face said OpenAI's two models — one publicly released and a second unreleased — did much more: They carried out 17,600 hacking actions on the internet between July 9 and July 13, during which time the models moved from their first foothold on the open internet to inside Hugging Face's servers . . . While the techniques detailed in Hugging Face's analysis were not beyond the reach of most skilled hackers, the AI company wrote that the two models were able to reconnoiter and expose holes in the company's layers of cyber defenses much faster than any human.[8]'

Anthropic's *Claude Mythos Preview* provides another notable example of these emerging capabilities. During controlled cybersecurity testing, the model autonomously identified and exploited previously unknown software vulnerabilities across every major operating system, every major web browser, and other widely used software.[9] Many of these vulnerabilities had remained undiscovered for years or even decades despite repeated review by experienced cybersecurity professionals. Among its discoveries was a previously unknown flaw in OpenBSD, a security-focused operating system, that had existed for 27 years before being identified by Mythos. Following OpenAI's disclosure of the Hugging Face incident, Anthropic conducted a retrospective review of its own cybersecurity evaluations and identified three separate incidents in which Claude models gained unauthorized access to the infrastructure of three organizations after inadvertently reaching the open internet through a misconfigured evaluation environment.[10] Recognizing the significance of these capabilities, Anthropic did not release Mythos to the general public. Instead, the company launched Project Glasswing, a collaborative initiative that initially provided early access to a select group of trusted companies and infrastructure organizations to help identify and remediate vulnerabilities in critical software before broader

---

[7] OpenAI, *OpenAI and Hugging Face partner to address security incident during model evaluation* (Jul. 21, 2026).
[8] Politico, *OpenAI's rogue models roamed the internet for 4 days and staged a second attack* (Jul. 28, 2026).
[9] AI Security Institute, *Our Evaluation of Claude Mythos Preview's Cyber Capabilities* (Apr. 13, 2026).
[10] Anthropic, *Investigating three real-world incidents in our cybersecurity evaluations (Jul. 30, 2026).*

deployment. These increasingly capable systems have raised important questions about how governments, industry, and researchers should evaluate these technologies, manage emerging risks, and strengthen cybersecurity preparedness.

CALIFORNIA'S PREPAREDNESS

As home to many of the world's leading AI developers and technology companies, the state is both driving the development of frontier AI systems and responsible for protecting one of the nation's largest digital ecosystems. This includes state government operations, critical infrastructure, businesses, and nearly 40 million residents who increasingly rely on secure and resilient digital services.

California has established a statewide cybersecurity framework to improve preparedness and coordinate responses to cyber threats. The California Cybersecurity Integration Center (Cal-CSIC), housed within the Governor's Office of Emergency Services (Cal OES), serves as the state's central cybersecurity coordination hub, bringing together state, federal, local, tribal, academic, and private-sector partners to share threat intelligence, coordinate incident response, and assess risks to critical infrastructure. Working alongside the California Department of Technology (CDT), these efforts seek to strengthen cyber resilience as cyber incidents continue to increase in both frequency and sophistication.

In addition to the operational cybersecurity infrastructure, California has begun to establish a statutory governance framework for frontier artificial intelligence through SB 53 (Wiener, Chapter 138, Statutes of 2025), the Transparency in Frontier Artificial Intelligence Act.[11] The law applies to developers with annual gross revenues in excess of $500 million, whose frontier AI models are trained using a quantity of computing power exceeding 10^26 integer or floating-point operations. The law requires these developers to establish and publicly disclose Frontier AI frameworks describing their safety and security practices, report safety incidents resulting in death, bodily injury, or more than $1 billion in damages to the California Office of Emergency Services (Cal OES), and provide whistleblower protections for employees who report potential violations.[12] SB 53 seeks to complement California's existing cybersecurity institutions by creating mechanisms for evaluating and monitoring the development of increasingly capable AI systems, while enabling state agencies to receive information about significant AI safety incidents that may have implications for public safety or critical infrastructure.

With the emergence of frontier AI, California faces the opportunity to leverage these technologies to strengthen cyber defense while preparing for possible challenges. By bringing together top experts from multiple sectors, this hearing tackles how California can responsibly lead in cybersecurity innovation while safeguarding millions of its people.

---

[11] S.B. 53, 2025–2026 Reg. Sess. (Cal. 2025) (Chapter 138, Statutes of 2025).
[12] Jon Iwry, *SB 53: What California's New AI Safety Law Means for Developers* (Nov. 14, 2025).