S&C MEMOS

Treasury Secretary and Federal Reserve Chair Warn Bank CEOs About Cybersecurity Risks Posed by Anthropic's New AI Model

Model's Capability to Identify and Exploit Critical Software Vulnerabilities Poses Risks for All Companies

April 15, 2026 | 6 min read |

Summary

On April 7, 2026, Treasury Secretary Scott Bessent and Federal Reserve Chair Jerome Powell convened an urgent, closed-door meeting with the CEOs of some of the nation's largest banks to discuss the cybersecurity risks posed by Anthropic's newly announced AI model, Claude Mythos Preview ("Mythos").[1] According to news reports, the meeting was held to ensure that the banks are aware of the serious cybersecurity risks Mythos and similar AI models under development pose, and are taking action to defend their systems.[2]

As Anthropic has disclosed, Mythos has demonstrated an unprecedented ability to identify critical, previously unknown (or "zero day") vulnerabilities in software, as well as methods of exploiting those vulnerabilities to enable malicious actors to compromise computer systems. Acknowledging that Mythos "could make cyberattacks of all kinds much more frequent and destructive, and empower adversaries of the United States and its allies," Anthropic has delayed the public release of Mythos in order to allow certain major technology and other companies to access the model and work together to identify and repair vulnerabilities and weaknesses in critical software infrastructure, an initiative that Anthropic has termed "Project Glasswing."

Notwithstanding the ongoing work of Project Glasswing, the sheer scale and interdependent nature of software and systems in global use means that it is likely that critical vulnerabilities that can be leveraged by bad actors using Mythos and similar AI models in the future will persist. Companies should take action now to ensure that they are minimizing the likelihood that exploitation of zero-day vulnerabilities in their systems will cause harm to them or others. This includes prioritizing risk management in their software supply chain, focusing on faster software patching processes, assessing their authentication and access controls, and investing in tailored threat-detection capabilities.

Notably, as the work of Project Glasswing illustrates, while Mythos and other AI models pose cybersecurity risk, they also create significant opportunities to enhance cybersecurity defenses and build software more securely. In announcing Mythos on April 7, Anthropic disclosed that it has been in ongoing discussions with the U.S. government not only about the model's offensive cyber capabilities, but about its defensive cyber capabilities. The meeting convened by Treasury Secretary Bessent and Federal Reserve Chair Powell with bank CEOs the same day was undoubtedly a reflection of those discussions. Companies should assess whether and how they can leverage leading AI models to enhance their cybersecurity going forward.

Background

Mythos is the newest and most advanced of Anthropic's general-purpose, large language AI models, which Anthropic has trained on a proprietary mix of publicly available information and private data sets. As described by Anthropic, during testing, the company discovered that Mythos "is capable of identifying and then exploiting zero-day vulnerabilities in every major operating system and every major web browser when directed by a user to do so."[3] Engineers "with no formal security training" asked Mythos to identify software vulnerabilities overnight that could be exploited by a third party acting remotely, and woke up in the morning to discover not only that Mythos had done so, but that the model had also provided them with a complete, fully functional method of exploiting those vulnerabilities.[4] Furthermore, as described by Anthropic, Mythos is capable of stringing together vulnerabilities that are not independently critical and using them to develop exploits to a degree that surpasses that of any existing AI model.[5]

Anthropic "did not explicitly train Mythos [] to have these capabilities," which instead "emerged as a downstream consequence of general improvements in code, reasoning, and autonomy" in the model.[6] According to Anthropic, Mythos has already identified thousands of previously unknown, high-severity vulnerabilities, less than 1% of which have been repaired to date.

In announcing Mythos, Anthropic simultaneously disclosed that it was delaying the model's public release in order to work with major technology and other companies in its "Project Glasswing" initiative, the goal of which is "to use Mythos [] to help secure the world's most critical software" before the model's offensive cyber capabilities can be exploited by malicious actors. Specifically, through Project Glasswing, Anthropic is providing limited access to about 40 organizations that build or manage critical software infrastructure "to find and fix vulnerabilities or weaknesses in their foundational systems—systems that represent a very large portion of the world's shared cyberattack surface."[7]

Anthropic also announced that it will report publicly within 90 days on what it has learned, what vulnerabilities have been fixed, and what improvements have been made that can be disclosed.[8] In addition, Anthropic announced a plan "to collaborate with leading security organizations to produce a set of practical recommendations for how security practices should evolve in the AI era."[9] This may include recommendations on vulnerability disclosure and software update processes; open-source and supply-chain security; software development lifecycle and secure-by-design practices; standards for regulated industries; and other matters.[10] Finally, Anthropic observed that cybersecurity in the age of powerful AI models will require joint work by the public and private sectors, and suggested that "an independent, third-party body—one that can bring together private- and public-sector organizations—might be the ideal home for continued work on these large-scale cybersecurity projects."[11]

Mythos has emerged in the context of an already rapidly escalating AI-enabled threat environment. According to the CrowdStrike 2026 Global Threat Report, AI-enabled attacks rose 89% year-over-year in 2025.[12] Recent research from the World Economic Forum found that 87% of organizations identified AI-related vulnerabilities as the fastest-growing cyber risk in 2025, more than any other category of threat, and a recent EY study reported that 96% of senior security leaders view AI-enabled cyberattacks as a significant threat.[13]
As Project Glasswing illustrates, however, AI models like Mythos are equally being leveraged to enhance cybersecurity defense, and may over time contribute as much or more to the enhancement of cybersecurity than to the creation of cybersecurity risk.

What Companies Can Do

In the near term, companies should expect that Project Glasswing will result in the identification of a greater number of critical vulnerabilities and corresponding software patches that companies will need to implement quickly to minimize the risk of compromise by malicious actors.

In addition, notwithstanding the work of Project Glasswing, companies should expect that Mythos and similar models will identify critical, zero-day vulnerabilities in software currently in use that malicious actors will be able to exploit before companies can patch them. While these vulnerabilities are by definition currently unknown and not reasonably discoverable by software users, there are important steps that companies can and should take now to minimize the likelihood of successful exploitation of zero-day vulnerabilities in their systems.

- *Patch management*. Companies should assess their patch management programs. Patch management is a critical, but often challenging, area in cybersecurity. Successful deployment of a software patch requires that the company first have a complete and up-to-date inventory of its technology devices and applications, that it have sufficient trained personnel to ensure adequate testing and timely deployment of patches, and that it adequately report and track its patch deployment to ensure compliance and identify gaps, among other things.
- *Supply chain risk*. Companies should assess risks in their software supply chain. Software sourced from smaller vendors with fewer or less sophisticated security resources could potentially be more vulnerable to AI-enabled compromise. Companies should, where feasible, evaluate the security posture of their vendors, require contractual commitments from vendors to secure software development and timely disclosure of cyber incidents, and consider reducing reliance on software from vendors that cannot demonstrate robust security practices.
- *Authentication and access controls*. Companies should ensure that their systems are appropriately segmented, and that access to their systems is appropriately limited, so that a compromise of one system does not cascade throughout the enterprise. In this regard, companies generally focus on implementing "zero-trust" security architecture principles, which treat every user, device and application as untrusted by default. The National Security Agency issued "Zero Trust Implementation Guidelines" in January 2026 to provide a structured, phased approach for companies to adopt Zero Trust cybersecurity frameworks. The guidelines emphasize continuous authentication and authorization of users, and operate under the principles of "never trust, always verify" and "assume breach."[14]
- *Detection capabilities*. The difference between a compromise that causes harm and one that does not often come down to timely detection of, and rapid response to, the threat. While it is essential for companies to invest in high-quality detection systems, doing so is not sufficient: companies need to ensure they have programmed these systems to provide the tailored, heuristic alerting that is essential for the timely identification of actual compromises. Companies should equally ensure that they have the resources and expertise to respond timely and effectively to any compromise that is identified.

None of these recommended measures is new, and all are among the best practices companies have long followed to mitigate cybersecurity risk. But as underscored by the magnitude of the cyber risk that Anthropic has described Mythos as posing, and the urgency of the meeting convened by Treasury Secretary Bessent and Federal Reserve Chair Powell to discuss this issue with CEOs of some of the nation's largest banks, it is critical for companies to devote heightened attention to these measures, and their cybersecurity programs more generally, at this time.

---

[1] *CNBC*, "Treasury Secretary Bessent, Fed Chair Powell met with bank CEOs to discuss Anthropic AI cyber risks," Apr. 8, 2026.

[2] *Financial Times*, "Scott Bessent called in US bank CEOs to discuss Anthropic model's cyber risks," Apr. 9, 2026; *NBC News*, "Treasury Secretary Bessent and Fed Chair Powell meet with bank CEOs about cybersecurity risks from new AI model," Apr. 8, 2026.

[3] Anthropic, "Assessing Claude Mythos Preview's cybersecurity capabilities," Apr. 7, 2026.

[4] *Id.*

[5] *Id.*

[6] *Id.*

[7] Anthropic, “Project Glasswing,” Apr. 7, 2026.

[8] *Id.*

[9] *Id.*

[10] *Id.*

[11] *Id.*

[12] *CrowdStrike*, “2026 Global Threat Report,” 2026.

[13] *World Economic Forum*, “Global Cybersecurity Outlook 2026,” 2026; EY, “Cybersecurity leaders investing in AI and agentic defenses to combat escalating AI-enabled threats,” Mar. 19, 2026.

[14] *NSA*, “Zero Trust Implementation Guideline Primer,” Jan. 2026.