Treasury Rushes To Access Anthropic 'Mythos' AI After Warning It Can Hack "Every Major Operating System"

April 14, 2026 By: Tyler Durden Zero Hedge

https://www.zerohedge.com/technology/treasury-rushes-access-anthropic-mythos-ai-after-warning-it-can-hack-every-major-system

The US Treasury Department's technology team is actively seeking access to Anthropic PBC's highly restricted Mythos AI model so it can begin hunting for software vulnerabilities, according to a person familiar with the situation cited by *Bloomberg*.

Treasury Chief Information Officer **Sam Corcos** briefed the department's cybersecurity team on the technology last week **and has directed efforts to gain access to the model "as soon as this week."**

The request comes days after **Treasury Secretary Scott Bessent and Federal Reserve Chair Jerome Powell summoned top Wall Street CEOs** to an urgent meeting at Treasury headquarters. Executives were warned that **Mythos and similar frontier AI models could usher in a new era of heightened cyber risk.** Anthropic itself has cautioned that the model may be capable of powering sophisticated cyberattacks unless companies proactively test it against their own systems and build defenses ahead of any wider release.

At the meeting, bank leaders were strongly urged to take the model seriously and use it internally to detect vulnerabilities.

What Is Mythos and Why the Restrictions?

Anthropic introduced Mythos (also referred to as Claude Mythos Preview) as part of its new **Project Glasswing initiative**. In internal testing, the model demonstrated extraordinary offensive cybersecurity capabilities: **it was able to identify and exploit vulnerabilities "in every major operating system and every major web browser when directed by a user to do so."** In one documented case, **it wrote a web browser exploit that successfully chained together four separate vulnerabilities.**

Project Glasswing brings together ***Amazon Web Services (AWS), Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, the Linux Foundation, Microsoft, NVIDIA, and Palo Alto Networks*** *to address growing concerns within the cybersecurity community that AI models*

are now capable of discovering and exploiting vulnerabilities at a faster pace than humans can keep up with.

...

According to the post on Anthropic's website, ***the model's strong agentic coding and reasoning skills enable it to uncover and exploit security flaws when directed by the user that have existed for years****, even decades without detection. Benchmarking results cited by the company suggest a notable performance gap between Mythos Preview and its previous models in cybersecurity-related tasks. -cxtoday.com*

What Mythos Has Discovered: Key Findings from Red Team Testing

In controlled testing against real codebases in isolated containers, the model autonomously identified**thousands of zero-day vulnerabilities across every major operating system** and **every major web browser**. The testing used an agentic workflow: file prioritization based on a 5-tier vulnerability likelihood ranking, parallel Claude Code invocations, and secondary validation for severity and exploitability.

Standout Zero-Day Discoveries Include:

- **27-year-old remote crash vulnerability in OpenBSD** (TCP SACK processing): An integer overflow in signed TCP sequence number comparison that enables a null-pointer dereference and remote denial-of-service against any responding host. The bug had survived decades of manual code review and extensive fuzzing campaigns.
- **16-year-old bug in FFmpeg** (H.264 parser): A slice number collision that triggers an out-of-bounds heap write when processing crafted frames with 65,536+ slices. The vulnerability originated in 2003, became exploitable after a 2010 refactor, and had evaded detection despite automated testing tools hitting the vulnerable path **five million times**.
- **17-year-old FreeBSD NFS Remote Code Execution (CVE-2026-4747)**: A stack buffer overflow in RPCSEC_GSS authentication (96-byte buffer for 304-byte input) combined with NFSv4 information disclosure. Mythos autonomously constructed a **20-gadget ROP chain** split across six sequential RPC requests — a feat the prior model (Claude Opus 4.6) could achieve only with significant human guidance.

Firefox JavaScript Engine Testing Results were especially dramatic:

- Claude Opus 4.6: Developed only **2** working exploits out of several hundred attempts.
- Mythos Preview: Developed **181** working exploits and achieved register control in **29** additional cases.

OSS-Fuzz Results showed a similar leap:

- Mythos generated 595 tier-1/2 crashes (plus several tier-3–5), including multiple **tier-5 control-flow hijacks** (full arbitrary code execution) on fully patched targets.

These discoveries were achieved at remarkably low cost - many individual zero-day runs cost under **$50**, with full OpenBSD testing campaigns under $20,000 and Linux kernel N-day exploits under $2,000 each.

Because of the dual-use risks, Anthropic has not released Mythos to the public. Instead, it is being provided on a tightly limited basis through Project Glasswing to a select group of vetted organizations - including **major tech companies, cybersecurity firms, JPMorgan Chase, and the Linux Foundation** - for defensive purposes only (scanning their own systems to find and patch flaws before attackers can exploit them). **Anthropic has committed up to $100 million in usage credits to support these efforts.**

Several major financial institutions have already begun internal testing:

- **JPMorgan Chase** was publicly named as part of Project Glasswing.
- **Goldman Sachs, Citigroup, Bank of America, and Morgan Stanley** have also gained access or are in the process, according to people familiar with the matter.

The company stated in its Project Glasswing announcement that it has been in "ongoing discussions" with government officials about the model and is "ready to work with local, state, and federal representatives."

Pentagon Supply-Chain Risk Designation

The Treasury's push for access is notable because the **Pentagon** formally designated Anthropic a US supply-chain risk earlier this year following a dispute over how the company's AI technology could be used by the military. The Defense Department gave Anthropic a six-month window to transition its services to another provider. Anthropic is actively fighting the designation in federal court.

Despite this, Corcos - who previously encouraged the use of Anthropic's Claude AI tools inside Treasury before the Pentagon label - is now driving the department's effort to investigate Mythos.