Former national cyber director: Anthropic's 'Mythos' AI can hack nearly anything and we aren't ready

April 23, 2026 By: Kemba Walden Fortune
https://tech.yahoo.com/cybersecurity/articles/former-national-cyber-director-anthropic-100000869.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAALEbETiGJyspDuww_dWV6zUTu84CxfdGCCnV2gAvf-Lj9yOoe5jD7Tvpc-QZm8HmliEUtHVar__e4Eb8FfmavxfObFPh_AVk06T8ExT-RzB1Akf9OohEWyC9pSELjASyPXLeFsPw9Pa1VbSAmhRhxDhM3j1k1MNVdQiYoY2Ih9kg

Kemba Walden served as Acting National Cyber Director of the United States and is President of the Paladin Global Institute. (courtesy of Kemba Walden)

Artificial Intelligence has evolved from a theoretical concept to a prevailing element of daily life. And while the generative AI boom has brought significant improvements to our society, Mythos, Anthropic's most advanced model, should be a clarion call to address weaknesses in our cyber ecosystem.

Anthropic's own cyber experts have called the model "too powerful to be released to the public" at this stage due to its sophisticated capabilities and ability to carry out advanced attacks. The model represents a leap in defensive AI capabilities, but it also possesses inherent risks that expose vulnerabilities in our critical infrastructure and

systems. Not only does the model discover zero-day vulnerabilities, but it autonomously builds and chains exploits—and then covers its tracks—making it more difficult to defend against them. Experts are rightfully asking: how will Mythos impact markets, and should we fear its potential in the hands of America's adversaries?

With the development of Mythos and the ability to exploit vulnerabilities, the technical debt in U.S. critical infrastructure is coming due. We've long known that U.S. critical infrastructure is owned and operated by private sector companies of all sizes. Compounding the risk, some functions like water systems and power distribution are run by under-resourced state and local agencies. While Fortune 500 enterprises are better equipped to update their infrastructure, the vast majority—small and medium enterprises (SMEs) and smaller agencies—are less empowered to make these upgrades. We must catalyze investing in these under-resourced corners of infrastructure so that newer technologies built to withstand AI-enabled exploitation can be effective.

Today, we need urgent investment, policy innovation, and public-private collaboration to ensure AI strengthens, rather than undermines, our national security.

Mythos is said to possess the ability to discover flaws in "virtually any and every operating system, browser, or other software product." The model has sparked concern among global financial institutions regarding how it could respond to the scale and rate of AI-powered exploits. Indeed, Mythos has already demonstrated an 83 percent success rate in exploit creation on the first attempt. It's not just that Mythos can discover vulnerabilities and autonomously build exploits for them—it's capable of chaining those exploits together, making the challenge of defending against them far greater. This isn't merely a technical feat, but a call to evolve our modern cybersecurity strategies as AI acts as both a tool and a threat multiplier.

For now, Mythos is operating under a limited release initiative to preview this version for industry partners like Microsoft, AWS, Google, and NVIDIA, who can identify flaws in the system before our adversaries are able to. Palo Alto Networks, a partner in Project Glasswing, has called the model a "game changer" in uncovering hidden defects. Anthropic has also briefed key U.S. officials including members of the Cybersecurity and Infrastructure Security Agency and the Center for AI Standards and Innovation regarding the model's sophisticated capabilities.

Anthropic's Mythos Preview is a wake-up call: AI's rapid development demands we invest in our infrastructure's weakest links and innovate to protect critical systems. Policymakers, cyber and AI experts, and developers must address this technical debt through research and advocacy for public-private partnerships that direct investment

dollars toward under-resourced sectors. By prioritizing incentives for AI-resilient technologies, the information technology sector can ensure that SMEs aren't left behind in the race against AI-driven threats.

Mythos's unexpected sandbox breakout—which bypassed its own security guardrails—has surprised many security researchers and sparked vital discussions. This incident, detailed in the preview's system card, offers a prime opportunity for technologists, industry leaders, and policymakers to collaborate on solutions to start building reliable security for these systems.

AI should bolster, not erode, our collective defenses. By fostering partnerships between government, sector leaders, and innovators, the U.S. can help deploy resilient technologies that not only withstand exploit chaining but possess long-term resilience.

The opinions expressed in Fortune.com commentary pieces are solely the views of their authors and do not necessarily reflect the opinions and beliefs of Fortune.

This story was originally featured on Fortune.com