A secretive AI hacking system has sparked a global scramble

April 24, 2026 By: Ian Duncan The Washington Post

https://www.washingtonpost.com/technology/2026/04/24/anthropic-mythos-ai-washington-cybersecurity-hacking-risk/

When security researchers at Mozilla, the maker of the popular web browser Firefox, pointed a powerful new artificial intelligence model at their code, they had a feeling of "vertigo."

Bobby Holley, the chief technology officer for the browser, said Anthropic's Mythos system elevated AI from being merely a competent software engineer to "a world-class, elite security engineer."

"It was just a 'wow' moment," Holley said in an interview.

Almost a hundred engineers set aside other work to begin tackling a wave of security problems. The latest version of Firefox contains fixes for 271 flaws found with Mythos's help. Any one of them would have been a red-alert moment just last year, Holley wrote in a blog post this week. The most serious vulnerabilities in older versions of the browser theoretically could be used to install programs or delete data, according to an advisory from the Center for Internet Security, although there is no evidence of them being put to use.

The findings were some of the first since Anthropic announced Mythos to the world this month, saying the new generative AI model could supercharge the abilities of hackers, making it too dangerous to release to the general public. The powers represent the shadow side of the company's popular Claude Code tool - its system's ability to almost miraculously pump out line after line of code, meaning it can also understand how to break it apart.

Computer security experts say they have long foreseen the day AI becomes a formidable hacking tool. But in Washington and foreign capitals alike, Anthropic's announcement has hit hard, setting off a scramble by government officials to understand what is at risk and reinvigorating a debate over what safety guardrails they should impose on powerful AI systems.

Yet it remains unclear just how significant of a threat Mythos and other advanced AI models will pose in the wild. The model has been limited to a small circle of businesses, wrapped in nondisclosure agreements. Sam Altman, the chief executive of OpenAI, suggested this week that his rivals at Anthropic were engaging in "fear-based marketing."

Holley's tests left him optimistic rather than completely terrified, with the tools providing the means to close security holes that were difficult to find before.

"I am really positive on the timeline that we are in right now, with the capabilities making their way into the hands of defenders first," he said.

Nonetheless, the development has added urgency to some of the concerns about AI, which have focused on remote fears of mass job losses or theories of rogue robots running amok. Mythos suggests more immediate problems, involving waves of hacked bank accounts or hospital computers locked up by criminals demanding ransoms.

“Mythos has activated a lot of people in D.C.,” said Dean Ball, a former White House AI adviser. “AI has become the top priority for a lot of people for whom it hasn’t been.”

For the Trump administration, which has close ties to the tech industry and has been bullish about the potential of AI to unlock an economic revolution, the arrival of Mythos has led to a reckoning with some of the technology’s potential downsides.

The White House has tasked its Office of the National Cyber Director with coordinating a response and is drawing on the expertise of the National Security Agency - home to the government’s own elite hackers - to get a handle on the danger, according to people briefed on the efforts. Anthropic chief executive Dario Amodei visited the White House last week to brief senior officials, even as his company remains locked in a legal fight with the government over use of its systems by the military.

A White House official said the Trump administration was “exploring the balance between advancing innovation and ensuring security,” adding that “the collective effort of all involved will ultimately benefit our country and economy.”

Brendan Steinhauser, the chief executive of the Alliance for Secure AI, said it was encouraging that the administration had responded quickly to the “Mythos moment.”

“I’m glad to see that the president and his administration take this issue very seriously and elevate this issue to the top of their priority list,” said Steinhauser, whose organization has clashed with some on President Donald Trump’s team over previous efforts by the administration to block AI regulations.

As the Mozilla results suggest, the newest AI models could prove adept at finding fresh security flaws in computer code - vulnerabilities known as “zero-days.” Anthropic says it found one that had lurked undetected for 27 years. This could allow companies and governments to reinforce their digital holdings. At the same time, experts say such tools could allow hackers to automate their attacks, speeding up their operations and making it possible for even people with no computer security training to stage digital break-ins.

Evan Peña, one of the founders of Armadin, a security company that uses AI to break into customers’ systems in order to find and fix flaws, said the models have become more capable faster than he expected.

“Now you can have 1,000 Evan Peñas constantly coming at you,” he said.

The model itself remains largely shrouded in mystery, despite the release of a colorful 245-page document outlining its development. (At one point, it successfully demonstrated that it could break free of restrictions and sent an “unexpected” email to a researcher while they were eating a sandwich in the park, according to the document.)

Anthropic formed a partnership dubbed Project Glasswing with a handful of leading tech companies and other big businesses so they could began assessing the risks to their own systems, but few findings have been released.

The results from outsiders that have emerged present a mixed picture. The British government's AI Security Institute assessment of Mythos found that while it was more capable than older systems on a battery of tests - succeeding 73 percent of the time on difficult tasks that no AI could complete until last year - how that would translate into real world dangers remained to be seen. The assessment said the AI Security Institute test environment did not have the kind of active defenses many fortified systems employ, handing the AI model an advantage in testing.

Anthropic has said it had discussions with federal cybersecurity agencies, including the Center for AI Standards and Innovation. The center did not respond to questions about whether it planned to release its own analysis of Mythos.

While the risks are relatively contained right now, other AI companies - including those overseas - are expected to develop their own tools with similar capabilities in coming months. A Switzerland-based security researcher reported this week that one Chinese firm appeared to already be employing techniques similar to those enabled by Mythos.

There is also the question of whether Anthropic will be able to ensure Mythos is used for good. The company confirmed this week that it was investigating a report by Bloomberg News that outsiders had gained access to the tool. The news service reported that a group organized on the chat app Discord made an educated guess about the model's online location and drew on one person's access to Anthropic's systems via a contractor. The group's goal was to play with the model, rather than use it for cybersecurity, according to the report.

OpenAI, Anthropic's chief rival, also said this month that it has developed a new version of its ChatGPT model that is adept at cybersecurity tasks. OpenAI said it had briefed federal cybersecurity agencies about the model and was expanding a program that will let approved developers use it to harden their systems, initially working with a mix of tech companies and big businesses that is similar to the Anthropic partnership. It held a meeting with dozens of federal cybersecurity experts this week to demo its latest technology and briefed officials at the White House and on Capitol Hill Thursday, according to a person familiar with the events. The meeting with cybersecurity experts was first reported by Axios.

But by pulling off its dramatic rollout of Mythos and Project Glasswing first, Anthropic again showed its ability to set the agenda for the AI industry. Several times in recent months the company has announced products that have rattled existing businesses. And by commanding the attention of governments, it has built the potential for a reset in its relationship with the White House.

The Trump administration had sought to push Anthropic aside this year, booting it from the Pentagon's systems and banning other federal agencies from working with it. Officials accused the company of strong-arming the government in a dispute over how what safeguards its model should have when they are used by the military.

“Their selfishness is putting AMERICAN LIVES at risk, our Troops in danger, and our National Security in JEOPARDY,” Trump wrote on his Truth Social site in February.

But this week, after Amodei’s visit to the White House, the president’s tone changed.

“We had some very good talks with them, and I think they’re shaping up,” Trump said Tuesday in an interview on CNBC. “They’re very smart, and I think they can be of great use.”