LILY HAY NEWMAN SECURITY JUN 16, 2026 1:50 PM

‘Dangerous’ AI Models Are Coming No Matter What

The US government crackdown on Anthropic’s Claude Fable 5 and Mythos 5 hides a glaring truth: AI models with advanced hacking capabilities will soon be the norm.

Anthropic CEO Dario Amodei. PHOTOGRAPH: JASON HENRY/BLOOMBERG/GETTY IMAGES

Listen • 5 minutes

LATE LAST WEEK, Anthropic took its new Claude Fable 5 and Mythos 5 AI models

House since Friday but has yet to secure an agreement that would allow it to reinstate the offerings.

Since Mythos debuted in April, Anthropic has claimed—and warned—that the model has advanced capabilities for not only finding software vulnerabilities to help defenders patch them, but also figuring out ways to exploit them that could be used by bad actors. Anthropic itself noted this double edged sword in its launch of Mythos 5 and Claude Fable 5. “A great deal of advanced usage of AI models is dual use: the same queries that are beneficial in the hands of cybersecurity professionals and biology researchers could be dangerous if available to malicious actors,” the company wrote in a blog post last week.

With this in mind, the company initially released a version called Mythos Preview to a select consortium as part of a working group known as Project Glasswing. Mythos 5 was also privately released to this group last week, while Claude Fable 5, which is a Mythos-grade model, was released to the general public with specific blocks on its ability to give responses to questions about biology and cybersecurity.

Then, at the end of last week, the Trump administration moved to restrict both models because it believes that Fable 5’s guardrails can be disabled to allow full access to the Mythos 5 capabilities, allegedly making it a national security risk.

Experts say, though, that this institutional clash is simply delaying or masking a hard truth: Anthropic may be the tip of the spear in this moment, but AI capabilities in general and models from multiple companies and open-weight developers will almost certainly have similar capabilities to Mythos 5 in the near future—if they don't already.

“It's myopic in the extreme to think that no other competitors to Anthropic will develop similar capabilities to Mythos or even that they have not already done so,” says Tarah Wheeler, chief security officer of the specialized cybersecurity consulting firm TPO Group. “There are other companies hot on Anthropic's heels who probably have the capabilities, too, and are holding them in reserve as they see how Anthropic is being treated in the current regulatory environment.”

Anthropic itself has emphasized this point since the launch of Mythos Preview. “The real message is that this is not about the model or Anthropic,” Logan Graham, the company's frontier red team lead, told WIRED when Mythos Preview launched in April. “We need to prepare now for a world where these capabilities are broadly available in 6, 12, 24 months.”

Researchers note that even before this next generation of models, existing AI offerings could be used for advanced vulnerability-hunting and exploit development with a refined harness. A large group of cybersecurity leaders emphasized this to the administration in an open letter on Sunday, arguing that the White House's export-control directive was misguided.

“It's not one model; it's the general trend of technology,” says Bruce Schneier, a researcher at Harvard University and the University of Toronto who has been analyzing the situation. “Smaller, cheaper, open-source models, sometimes by themselves and sometimes in concert with each other, can match Mythos/Fable's performance with more sophisticated prompting. And we should expect other models to match Mythos/Fable's creativity and tenaciousness within months—slightly longer for open-source models.”

What the White House and governments around the world need to focus on, experts say, is democratically developing much broader and more transparent plans for how they will contend with advances in AI capabilities on cybersecurity and in other sensitive areas as they inevitably occur.

“The policy question is not whether a technology has risk,” says Chris Wysopal, cofounder of the cloud security firm Veracode. “The question is whether a specific restriction meaningfully reduces that risk or whether it mainly slows down the people trying to make systems safer.”

You Might Also Like

Lily Hay Newman is a senior writer at WIRED focused on information security, digital privacy, and hacking. She previously worked as a technology reporter at Slate, and was the staff writer for Future Tense, a publication and partnership between Slate, the New America Foundation, and Arizona State University. Her work ... Read More

SENIOR WRITER

TOPICS CYBERSECURITY ARTIFICIAL INTELLIGENCE HACKING SECURITY VULNERABILITIES ANTHROPIC OPENAI