Is Anthropic’s Claude Mythos Really a Cybersecurity Risk?

Cade Metz, Kate Conger

Is Anthropic’s New A.I. Really That Scary? It Depends Whom You Ask.

Anthropic said that Claude Mythos was too dangerous to release to the public. That claim has reopened an old debate over cybersecurity.

Listen · 8:20 min

Anthropic shook up the tech world with the limited release of its Claude Mythos A.I. model.Credit...Jason Henry for The New York Times

May 12, 2026

The artificial intelligence company Anthropic said last month that it would limit the release of its latest A.I. system to a small number of organizations, including a handful of big tech companies like Microsoft and Google and groups that manage important pieces of the internet.

Called Claude Mythos, the new system was too powerful to share with the general public, Anthropic said, because hackers could use it to exploit security holes in computer networks with stunning speed.

Executives in Silicon Valley and officials in Washington were alarmed by what Mythos could do, and its release may have helped shake the Trump administration from its defense of A.I. from government regulation.

The White House is now considering government oversight over new A.I. models, through an executive order that would create an A.I. working group of tech executives and government officials to examine potential oversight procedures. Among the possible plans is a formal government review process for new A.I. models.

But more than a month after Mythos was released, cybersecurity experts still disagree on whether Anthropic made the right call. Some applaud the company for restricting who got their hands on Mythos. Others criticize Anthropic for not

sharing it with a wider pool of researchers who could try it and get a handle on what it can and cannot do. So far, it seems, the only consensus is there is no consensus about Mythos.

Anthropic shared the technology with about 40 organizations that maintain critical computer infrastructure, so that they could use the system to patch security vulnerabilities before hackers exploited them.

Only a handful of the groups or companies that have spent time using Mythos would discuss it with The New York Times. But companies and researchers that did not have access were happy to offer their thoughts on the way that Anthropic released its new A.I.

Their feedback so far has ranged from serious concern to a shrug. It could be some time before the broader tech community concludes whether Anthropic was right to limit the Mythos release — a challenge that Anthropic executives acknowledge.

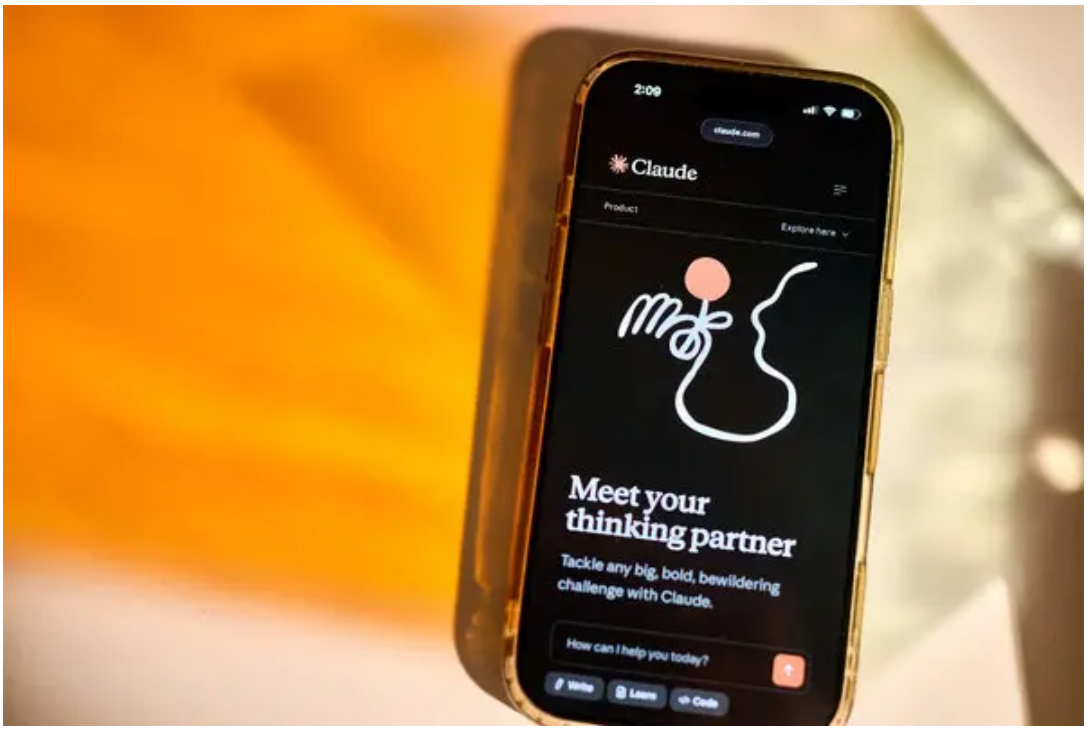
“For capabilities like this — or for a model as powerful as this — this is kind of an unprecedented situation where we truly do not have all the answers,” Logan Graham, head of Anthropic’s Frontier Red Team, which evaluates Claude for risks, said in an interview. “We don’t truly know what is the best way to roll out models like this.”

Experts can look at the same situation and come to very different conclusions because of the inherently complex nature of cybersecurity. People can use systems like Mythos to attack computer networks, but they can also use them to defend attacks. For decades, people have argued over the best ways of handling that dual nature.

Most experts agree that A.I. technologies like Mythos are fundamentally changing cybersecurity. The change gathered steam about six months ago when Anthropic and its chief rival, OpenAI, released new systems that are particularly good at writing computer code. If an A.I. system can write code, it can potentially find and exploit vulnerabilities in software applications.

When Anthropic unveiled Mythos, the company said it had used the technology to find thousands of security vulnerabilities that had gone undetected in popular software systems for years. Anthropic also said that Mythos was better at identifying disparate security flaws and stringing them together into “exploit chains,” which are used by malicious hackers to exploit several security holes as part of a coordinated attack. In the company’s words, the technology represented a “step change” in what was possible with A.I.

Image



Most experts agree that A.I. technologies like Mythos are fundamentally changing cybersecurity.Credit...Gabby Jones/Bloomberg

Cisco, the computer hardware and software company, is one of the companies that have used Mythos. Anthony Grieco, the company’s senior vice president and chief security and trust officer, said the technology is significantly more powerful than existing systems in certain areas.

Companies like Cisco, he said, should be “super aggressive about how we use this technology to identify vulnerabilities, fix them and get those fixes in the hands of our customers as rapidly as possible.”

He said that Mythos was indeed better at identifying exploit chains. But he added that those skills could be used to defend a computer network, not just attack it. “We are using that capability to help triage vulnerabilities and understand which ones are important to fix, so that sort of capability has a really positive connotation in the context of defense as well,” he said.

That is exactly why some cybersecurity researchers argue that Anthropic should release its system more widely. Like any other cybersecurity tool, it is good for both offense and defense.

“The technology is not too dangerous to release,” said Gary McGraw, a veteran security and A.I. researcher. “If you don’t release a tool like this — or you hoard it — you are not solving the real problem.”

Soon after Anthropic’s announcement, independent researchers showed that existing A.I. systems could find the same security holes that Mythos had found. Some cybersecurity experts argued that Anthropic had exaggerated the dangers of Mythos.

For Pavel Gurvich, co-founder and chief executive of the security company Tenzai, part of the problem is that independent cybersecurity experts are unable to test the system and gain a complete understanding of its strengths and weaknesses. That understanding can help them defend against attacks from the technology.

“I don’t think that choosing to share the model with such a small subset of companies helps us move forward,” Mr. Gurvich said. “This is especially true because the announcement was accompanied by very bold claims that we can’t assess.”

A week after Anthropic unveiled Mythos, its competitor OpenAI said that it, too, was sharing a similar technology only with a group of partners. But the company shared its model, GPT-5.4-Cyber, with a much larger group. It said it would initially share the model with hundreds of organizations, and then release it to thousands more partners in the coming weeks.

(The Times sued OpenAI and Microsoft in 2023 for copyright infringement of news content related to A.I. systems. The two companies have denied those claims.)

Mr. Gurvich said that this approach “made more sense,” in part because OpenAI has said that as it shares its technology, it will work to verify the identity of users in an effort to prevent misuse.

Stanislav Fort, a former Anthropic researcher who now runs a security company called Aisle, said that keeping A.I. technology bottled up will not be possible in the long run, because so many tech giants, start-ups and independent developers are building powerful systems. Many of these organizations are “open sourcing” their A.I., allowing anyone to use and modify the underlying technology.

As time goes on, he added, widely sharing these technologies will be essential to cybersecurity.

“Security by obscurity is one of the oldest bad ideas in the field,” he said.

Cade Metz is a Times reporter who writes about artificial intelligence, driverless cars, robotics, virtual reality and other emerging areas of technology.