Insights, News & Events

6 CYBERSECURITY STEPS YOU SHOULD TAKE AFTER ANTHROPIC'S CLAUDE MYTHOS OFFERS GLIMPSE INTO NEW WORLD OF AI DANGER

Insights
Apr 14, 2026

6 Cybersecurity Steps You Should Take After Anthropic's Claude Mythos Offers Glimpse Into New World of AI Danger

Anthropic's new AI model could enable such crippling damage in the wrong hands that the company has decided not to publicly release it – but it still reshapes the cybersecurity risk landscape for every business. The April 7 announcement of Claude Mythos sent shockwaves through the cyber community, and for good reason: the program already identified thousands of vulnerabilities across every major operating system and web browser, some of which had gone undetected for decades. At the same time, the company announced that over 50 technology and cybersecurity organizations would use Mythos to find and patch vulnerabilities in the world's most critical software products before adversaries can exploit them. Here's what businesses need to understand about this development and six steps you should take to get your cybersecurity house in order.

What is Claude Mythos, and how is it different from other AI models?

Claude Mythos sits well above Anthropic's current models and incorporates advanced agentic AI capabilities. It can autonomously plan, execute, and chain together complex multi-step tasks without human intervention.

Related People

Daniel Pepper, CIPP/US
Partner

303.218.3661

David J. Walton, AIGP, CIPP/US
Partner

Its cybersecurity capabilities are what distinguish it from every prior model. In pre-release testing, Anthropic found that Mythos Preview could not only identify undiscovered software vulnerabilities but also weaponize them. The model has already discovered thousands of high-severity zero-day vulnerabilities across every major operating system and web browser, including a 27-year-old bug in OpenBSD, a 16-year-old flaw in FFmpeg, and a memory-corrupting vulnerability in a memory-safe virtual machine monitor.

Essentially, this announcement means that AI has now crossed a threshold where it can outperform all but the most elite human hackers at finding and exploiting software vulnerabilities.

What is Project Glasswing?

Project Glasswing is Anthropic's initiative where over 50 companies (including Amazon Web Services, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorgan Chase, Microsoft, and Nvidia) will use Mythos Preview for defensive security work and share their findings with the wider industry. Anthropic is backing the project with $100 million in usage credits and has collected $4 million in direct donations to open-source security organizations.

The explicit purpose is to get ahead of the offensive use of Mythos-class tools before those tools reach adversaries. Anthropic has stated that the work of defending the world's cyber infrastructure may take years, while frontier AI capabilities are likely to advance substantially over just the next few months.

Why did Anthropic choose not to release Mythos publicly?

Citing the potential damage that could result from a wider public release, Anthropic released Mythos Preview only to a limited group of tech companies. It is the first time in nearly seven years that a leading AI company has so publicly withheld a model over safety concerns.

The specific concern is that Mythos enables autonomous, agentic cyberattacks at a scale and speed that defenders cannot currently match. Adversaries are already moving in this direction, and CrowdStrike's 2026 Global Threat Report found an 89% increase in attacks by adversaries using AI year-over-year.

610.230.6105

Service Focus

AI, Data, and Analytics

Data Protection and Cybersecurity

Privacy and Cyber

Resource Hubs

AI Governance Hub

Has AI already been used to conduct cyberattacks against real organizations?

Yes, and the documented cases are directly relevant to how businesses should be assessing their current risk exposure.

Anthropic has documented a case in which a Chinese state-sponsored group ran a coordinated campaign using Claude Code to infiltrate roughly 30 organizations – including tech companies, financial institutions, and government agencies, before the company detected it. Over the following 10 days, Anthropic investigated the full scope of the operation, banned the accounts involved, and notified affected organizations.

In November 2025, Anthropic blogged about what it described as the first reported AI-orchestrated cyber espionage campaign, in which attackers used AI's agentic capabilities not just as an advisor, but to execute the cyberattacks themselves.

What should businesses be doing right now?

Here are six steps you can take now to focus your efforts in light of this significant development:

1. Review your incident response plan for AI-attack scenarios. Most plans were not written with agentic AI attacks in mind. Map the specific scenario of an AI-orchestrated intrusion into your escalation workflows and pressure-test whether your team can detect, assess, and report within the 72-hour reporting window that will soon be required by federal CIRCIA rules.

2. Re-scope your third-party risk program. Vendor security addenda and questionnaires built around conventional patch timelines need to be updated. Mythos-class tools identify and exploit vulnerabilities in hours. "Reasonable" patching cadence language in your contracts may not be enforceable against that standard.

3. Audit your AI governance documentation. Board-level AI governance frameworks that focus only on responsible use of tools your organization deploys are incomplete. Governance programs should now explicitly address the risk that your AI deployments (or your vendors') can be manipulated or exploited by adversaries using comparable technology.

4. Review your cyber insurance policy terms. Confirm whether your policy language covers AI-orchestrated intrusions, autonomous exploit execution, and compressed attack timelines. Identify coverage gaps before an incident makes them consequential.

5. Review technology contracts governing AI tool deployments. Agreements for AI coding assistants, security platforms, and agentic workflow tools negotiated before mid-2025 were written without the Mythos risk framework in mind. You should review indemnification provisions, security warranty language, and liability caps against the new baseline.

6. Evaluate your detection and monitoring capabilities. You cannot report what you cannot see, and you cannot see an AI-accelerated attack with monitoring infrastructure designed for human-paced intrusions. Assess whether your current detection capability can identify the kind of lateral movement Mythos-class tools generate.

Conclusion

Fisher Phillips will continue to monitor developments and provide updates as warranted, so make sure you are subscribed to Fisher Phillips' Insight System to get the most up-to-date information direct to your inbox. If you have questions, contact your Fisher Phillips attorney, the author of this Insight, or any attorney on our Data Protection and Cybersecurity Team.