

Anthropic's Mythos AI model sparks fears of turbocharged hacking



AI-enabled cyber attacks were up 89 percent in 2025 compared with a year earlier, according to data from security group CrowdStrike. Meanwhile, the average time between an attacker first gaining access to a system and acting maliciously fell to 29 minutes last year, a 65 percent acceleration from 2024.

“The game is asymmetric; it is easier to identify and exploit than to patch everything in time,” said one person close to a frontier AI lab.

Anthropic's Graham said there were also internal concerns that companies would use Mythos to find “more vulnerabilities than they could hope to deal with in the near future.”

The heightened fears about AI and cyber security come amid signs that agents, which act autonomously on users' behalf to conduct tasks, could also fuel a further rise in AI-enabled

hacking.

Last September, Anthropic detected the first reported AI cyber-espionage campaign believed to be coordinated by a Chinese state-sponsored group.

It manipulated its coding product, Claude Code, to attempt to infiltrate about 30 global targets, including large tech firms, financial institutions, chemical manufacturers, and government agencies. It was successful in a small number of cases and executed without extensive human intervention.

Software researcher Simon Willison has warned there is a “lethal trifecta” of capabilities that arise with agents: access to private data; exposure to untrusted content, such as the Internet; and the ability to communicate externally.

Security professionals argue that the safest way to protect against cyber attacks when using an AI agent is to grant it access to only two of these areas. However, AI experts believe that much of the value from agents comes from granting access to all three.

“The bad news is that there is no good solution as of today,” said one person close to an AI lab. “The good news is [AI agents aren’t] yet in mission-critical settings like the stock exchange, bank ledger, or the airport.”

Stanislav Fort, a former Anthropic and Google DeepMind researcher who has founded AISLE, an AI security platform, said he was optimistic that AI could help to identify and fix a “finite repository” of historical security flaws.

To date, AI models have identified thousands of “zero-day” vulnerabilities—unknown weaknesses in commonly used software—some of which have been undetected for decades.

“We are gradually finding fewer and fewer zero days, of the worst kinds we can imagine,” said Fort.

Once these weaknesses were eliminated, the technology could be used to “proactively make sure nothing bad comes in [and] meaningfully increase the security level of the whole world as a result.”

Additional reporting by Kieran Smith in London.

© 2026 The Financial Times Ltd. All rights reserved. Not to be redistributed, copied, or modified in any way.

<https://arstechnica.com/ai/2026/04/anthropics-mythos-ai-model-sparks-fears-of-turbocharged-hacking/>