

AI-Powered Vulnerability Exploitation

On June 9, 2026, Anthropic released Claude Fable 5 to the public. Fable 5 is built on the same underlying model as Mythos, the controversial AI that until this week was restricted to a small number of approved organizations.

In recent weeks, media outlets worldwide reported on the unprecedented threat that Mythos-class AI poses. For those who may have missed that coverage, an excerpt from ZenData's April 13 report is provided below.

> ZENDATA
>
> *April 13 2026 -*
>
> *Claude Mythos found thousands of critical vulnerabilities in just a few weeks, some of them 27 years old, with zero human intervention. Its first-attempt exploitation success rate: 83%. Which raises an uncomfortable question for our industry, because Mythos is not a story about AI serving cybersecurity. It is the signal that the fundamental model on which the entire industry has built its defenses for thirty years has just become obsolete.*

Mythos was originally restricted because of its unparalleled ability to autonomously discover and exploit previously unknown software vulnerabilities with no human involvement. Its capabilities prompted emergency meetings between the US Treasury, the Federal Reserve, and major bank CEOs, and in May the IMF issued a formal public warning that Mythos-class AI poses serious risks to the global financial system.

Anthropic states that Fable 5 includes safety guardrails that block high-risk requests and route them to a less capable model. Although this may sound reassuring, every major AI model released to the public has been jailbroken, many in record time. In the context of AI, "jailbroken" refers to the various means by which threat actors trick or alter a model to circumvent or remove its guardrails.

The difference this time is that threat actors will have a tool capable of finding and weaponizing software vulnerabilities at machine speed. The window between a vulnerability being discovered and being used in attacks has shrunk dramatically. For this reason, we are recommending that critical and high-severity patches be applied within 72 hours of release, and that automated patching be adopted wherever practical.

We have also advised that strong access controls, modern endpoint detection, AI-driven threat hunting backed by a 24/7 human SOC, and vulnerability scanning and remediation be adopted across the board. With the release of Fable 5, any organization that has not adopted these measures is exposed in ways that seemed implausible even six months ago.

We recognize that an increased patching cadence and additional threat mitigation technologies represent a real financial burden for some organizations, on top of an already considerable stack. Given the severity of this new landscape, we do not want cost to be the reason any client remains exposed. For organizations where it makes sense, transitioning from a time-based engagement to an inclusive flat fee arrangement can make adopting the full set of recommended protections far more manageable, and we are glad to walk through what that would look like during your review.

Ref. Claude Fable 5 / Mythos AI Capability Advisory, June 2026

Further information is available via the links provided below.

Anthropic: Claude Fable 5 and Claude Mythos 5 Announcement

CSO Online: Anthropic Releases Mythos-Class Fable 5 with Safeguards for Cyber Risks

Bain & Company: Claude Mythos and the AI Cybersecurity Wake-Up Call

Darktrace: Defending in an Era of AI-Accelerated Vulnerability Discovery

Sincerely,

Chris Harper

Chief Security Officer

Sandbox Technologies, Inc.