Unauthorized Group Gains Access to Anthropic's Exclusive Cyber Tool Mythos

April 22, 2026 By: Guru Baran Cyber Security News

https://cybersecuritynews.com/anthropic-mythos-access/

A group of unauthorized users has reportedly breached access controls surrounding Claude Mythos Preview, Anthropic's powerful and closely guarded AI-driven cybersecurity tool, raising serious concerns about third-party vendor security and the risks of placing advanced offensive AI capabilities in the wrong hands.

Announced on April 7, 2026, Claude Mythos Preview is an AI model that Anthropic itself described as too dangerous to release publicly. The model, deployed under Anthropic's Project Glasswing initiative, is capable of discovering zero-day vulnerabilities across major operating systems and web browsers, chaining software bugs into multi-step exploits, a feat previously achievable only by the most skilled human hackers.

In one alarming pre-release evaluation, Mythos autonomously escaped a secured sandbox environment, devised a multi-step exploit to gain internet access, and even emailed a researcher all without being instructed to do so.

Given these capabilities, Anthropic restricted access to a curated consortium of over 40 elite technology companies, including Apple, Amazon, Microsoft, Google, NVIDIA, Cisco, and CrowdStrike, for the exclusive purpose of identifying and patching critical software vulnerabilities before hostile actors could exploit the same techniques.

Unauthorized Group Gains Access

Despite these precautions, Bloomberg News reported on April 21, 2026, that a small group of unauthorized users gained access to Mythos through a third-party vendor environment on the very same day the model was publicly announced.

The group, communicating through a private Discord channel dedicated to gathering intelligence on unreleased AI models, reportedly made an educated guess about the model's online location based on familiarity with Anthropic's URL formatting conventions for other models.

The breach was facilitated, at least in part, by an individual currently employed at a third-party contractor working with Anthropic.

Bloomberg reported that partners were granted access for penetration testing, and unauthorized users exploited shared accounts and API keys belonging to authorized contractors.

The unauthorized group has been regularly using Mythos since gaining access and has provided Bloomberg with proof in the form of screenshots and a live demonstration of the software.

The source reportedly described the group's intent as curiosity-driven, "interested in playing around with new models, not wreaking havoc" — though security experts stress that intent is irrelevant when the tool in question is capable of devastating cyberattacks.

Anthropic confirmed awareness of the situation in a statement to TechCrunch: *"We're investigating a report claiming unauthorized access to Claude Mythos Preview through one of our third-party vendor environments."*

The company added that, as of now, there is no evidence that the unauthorized access has impacted Anthropic's core systems or extended beyond the vendor environment.