ZEROFOX INTELLIGENCE

Flash

Pending Mythos Release Presents Unique Security Concerns

F-2026-06-04a

Classification: TLP:CLEAR

Criticality: Low

Intelligence Requirements: Artificial Intelligence, Threat Actor, Malware

June 4, 2026

Scope Note

ZeroFox Intelligence is derived from a variety of sources, including—but not limited to—curated open-source accesses, vetted social media, proprietary data sources, and direct access to threat actors and groups through covert communication channels. Information relied upon to complete any report cannot always be independently verified. As such, ZeroFox applies rigorous analytic standards and tradecraft in accordance with best practices and includes caveat language and source citations to clearly identify the veracity of our Intelligence reporting and substantiate our assessments and recommendations. All sources used in this particular Intelligence product were identified prior to 8:00 AM (EDT) on June 4, 2026; per cyber hygiene best practices, caution is advised when clicking on any third-party links.

| Flash | Pending Mythos Release Presents Unique Security Concerns

| Key Findings

- On May 28, 2026, artificial intelligence (AI) developer Anthropic announced the upcoming release in several weeks of its cybersecurity-focused model, Mythos. Mythos was first announced in April 2026 as a general purpose AI model with powerful capabilities for use in cybersecurity applications.
- Mythos likely marks a leap forward in AI capabilities that threat actors will endeavor to exploit in order to conduct successful ransomware and digital extortion (R&DE) attacks. Mythos Preview was able to autonomously conduct basic attack processes 30 percent of the time—and completed an average of 22 of 32 required steps for a successful attack.
- Threat actors have almost certainly used AI models since as early as 2023 to conduct attacks, bypassing guardrails put in place by developers—called "jailbreaking"—to get the AI models to conduct malicious activities outside the intended parameters.
- Some technological capability and knowledge would likely still be required to fully exploit Mythos once it becomes publicly available. The technological know-how

needed to effectively deploy Mythos as an R&DE tool will likely limit the actors capable of exploiting it in the near term to established R&D threat collectives.

Details

On May 28, 2026, AI developer Anthropic announced the upcoming release of its cybersecurity-focused model, Mythos, in "the coming weeks."[1] The announcement was made in a blog post on the company's web site and also included updates about different Anthropic models, including the release of Opus 4.8.

Mythos was first announced in April 2026 as a general purpose AI model with powerful capabilities for use in cybersecurity applications.[2] The initial version, dubbed Mythos Preview, was reportedly capable of identifying zero-day vulnerabilities in every major operating system and every major web browser when directed by a user to do so.

- In response to the unprecedented capabilities of Mythos, Anthropic initiated Project Glasswing: a multi-faceted initiative to utilize Mythos' capabilities to identify and close cybersecurity vulnerabilities before threat actors can use the power of AI coding to exploit them.[3]
- Participants in Project Glasswing include Amazon Web Services, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, the Linux Foundation, Microsoft, NVIDIA, and Palo Alto Networks.

Mythos likely marks a leap forward in AI capabilities that can be exploited to conduct successful R&DE attacks. According to the UK government's AI Security Institute, Mythos Preview was able to autonomously conduct basic attack processes 30 percent of the time—and across all test attempts, the model completed an average of 22 of 32 required steps for a successful attack.[4]

Additionally, internal testing by Anthropic resulted in Opus 4.6, a highly capable AI model within cybersecurity applications that proved unable to consistently develop exploits

[1] hXXps://www.anthropic[.]com/news/claude-opus-4-8
[2] hXXps://red.anthropic[.]com/2026/mythos-preview/
[3] hXXps://www.anthropic[.]com/glasswing
[4] hXXps://www.aisi[.]gov[.]uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities

autonomously in a Firefox web browser benchmark version. By contrast, Mythos Preview was able to autonomously develop 181 exploits in the same tests and achieved register control in 29 more.[5]

Testing on Mythos Preview demonstrates a highly capable model that can very likely be used by threat actors maliciously to significantly improve AI-assisted R&DE attacks. However, at present, Mythos is almost certainly limited in its ability to successfully attack systems with robust, layered defenses; alternately, less well-defended systems will likely be even more vulnerable with the public release of Mythos.

The Ransomware Kill Chain Under AI Compression

In September 2025, Anthropic disclosed that a state-aligned threat actor used an AI coding agent to execute a sophisticated, automated cyber espionage campaign.[6] AI handled 80–90 percent of attack operations (including writing exploit code, performing reconnaissance, and attempting lateral movement at machine speed) against 30 targets across the globe. This disclosure presents an emerging security threat—one in which threat actors no longer need weeks or months to go through the process of an attack but can instead maliciously use AI agents to reduce the attack timeline, very likely to hours.

The cyber kill chain, as it existed before AI, assumed that human threat actors would have to go through laborious steps to perform reconnaissance, gain access, move laterally in a system, maintain persistence, escalate privileges, and exfiltrate data. However, AI almost certainly compresses this kill chain timeline down to hours, and Mythos very likely represents a security challenge that organizations should treat as a current reality.

- The stages of an attack where defenders currently have the most detection and mitigation capabilities—lateral movement, reconnaissance, and exfiltration—are very likely the stages where threat actors can use Mythos-class models to compress the kill chain, creating a strategic risk for security teams.

[5] hXXps://red.anthropic[.]com/2026/mythos-preview/
[6] hXXps://thehackernews[.]com/2026/03/the-kill-chain-is-obsolete-when-your-ai.html

Mythos Restrictions Are a Delay, Not a Barrier

As early as 2023, threat actors have almost certainly used AI models to conduct attacks, bypassing guardrails put in place by developers (known as "jailbreaking") to instruct AI models to conduct malicious activities outside the intended parameters. In early 2023, actors developed WormGPT, a malicious generative AI model designed to facilitate business email compromise (BEC) and phishing campaigns.[7]

WormGPT operated as a purpose-built large language model (LLM) designed to facilitate crime by bypassing security guardrails in mainstream AI models such ChatGPT and Google Gemini, spawning a new type of commercialized, criminal AI model called a "Dark LLM."[8]

Anthropic has almost certainly put guardrails in place for Mythos, which it likely believes will limit threat actors' ability to use the model for malicious ends. Additionally, it is very likely Anthropic will continue to regularly patch Mythos post-public release to further strengthen the full version against threat actor exploitation. However, threat actors will almost certainly use shared knowledge in the underground, criminal ecosystem to jailbreak subsequent versions. This will very likely create a continuously opening-and-closing guardrail window, wherein Anthropic sprints to patch Mythos before threat actors can further jailbreak the model.

In April 2026, members of an online cybersecurity forum were able to gain access to Mythos Preview without permission.[9] It is unlikely that threat actors gained access to the model; however, the unauthorized access to Mythos Preview stands as a demonstration, and very likely a warning, about the likely weakness of AI security—a reality made even more stark by Mythos' capabilities.

While AI coding agents almost certainly lower the technological bar to entry for less sophisticated threat actors, it is unlikely that inexperienced actors will adopt and benefit from Mythos at public launch. Although Mythos performed exceptionally well in capture-the-flag tests, it very likely required human guidance to complete the steps

[7] hXXps://www.huntress[.]com/cybersecurity-101/topic/wormgpt
[8] *Ibid.*
[9] hXXps://www.bbc[.]co[.]uk/news/articles/cy41zejp9pko

required to conduct a successful attack. This suggests that exploiting Mythos for cybercriminal purposes would likely still require some technological capability and knowledge—attributes established R&DE collectives possess.

The public release of Mythos very likely represents a significant elevation in the threat from AI-assisted attacks. However, the primary threat is likely to networks and systems with insufficient defenses. Networks with robust, layered, and active cyber defenses, while still likely to be targeted, will likely have a higher degree of success in disrupting Mythos-assisted attacks.

Security fears regarding the use of AI to lower the barrier to entry for less experienced actors are well placed. However, in light of the technological know-how required to effectively deploy Mythos as an R&DE tool once it is publicly available, actors capable of exploiting it for malicious purposes will likely be limited to established R&D threat collectives in the near term.

Long term, AI will almost certainly lead to an increase in overall attacks across all business sectors. AI coding agents, especially powerful models such as Mythos, very likely allow knowledgeable actors to conduct all steps required for a successful attack in a much shorter amount of time; however, the amount of time it will take to provide security patches to protect a network is also very likely to be compressed.

Recommendations

- Develop a comprehensive incident response strategy.
- Deploy a holistic patch management process, and ensure all IT assets are patched with the latest software updates as quickly as possible.
- Adopt a Zero-Trust cybersecurity architecture based upon a principle of least privilege.
- Implement network segmentation to separate resources by sensitivity and/or function.
- Ensure critical, proprietary, or sensitive data is always backed up to secure, off-site, or cloud servers at least once per year—and ideally more frequently.
- Implement secure password policies, phishing-resistant multi-factor authentication (MFA), and unique credentials.
- Configure email servers to block emails with malicious indicators, and deploy authentication protocols to prevent spoofed emails.
- Proactively monitor for compromised accounts and credentials being brokered in deep and dark web (DDW) forums.
- Leverage cyber threat intelligence to inform the detection of relevant cyber threats and associated tactics, techniques, and procedures (TTPs).

Appendix A: Traffic Light Protocol for Information Dissemination

	Red	Amber
WHEN SHOULD IT BE USED?	**Sources may use** TLP:RED when information cannot be effectively acted upon by additional parties and could lead to impacts on a party's privacy, reputation, or operations if misused.	**Sources may use** TLP:AMBER when information requires support to be effectively acted upon but carries risks to privacy, reputation, or operations if shared outside of the organizations involved.
HOW MAY IT BE SHARED?	**Recipients may NOT share** TLP:RED with any parties outside of the specific exchange, meeting, or conversation in which it is originally disclosed.	**Recipients may ONLY share** TLP:AMBER information with members of their own organization and its clients, but only on a need-to-know basis to protect their organization and its clients and prevent further harm. **Note that** TLP:AMBER+STRICT restricts sharing to the organization only.

	Green	Clear
WHEN SHOULD IT BE USED?	**Sources may use** TLP:GREEN when information is useful for the awareness of all participating organizations, as well as with peers within the broader community or sector.	**Sources may use** TLP:CLEAR when information carries minimal or no risk of misuse in accordance with applicable rules and procedures for public release.
HOW MAY IT BE SHARED?	**Recipients may share** TLP:GREEN information with peers and partner organizations within their sector or community but not via publicly accessible channels.	**Recipients may share** TLP:CLEAR information without restriction, subject to copyright controls.

Appendix B: ZeroFox Intelligence Probability Scale

All ZeroFox intelligence products leverage probabilistic assessment language in analytic judgments. Qualitative statements used in these judgments refer to associated probability ranges, which state the likelihood of occurrence of an event or development. Ranges are used to avoid a false impression of accuracy. This scale is a standard that aligns with how readers should interpret such terms.

Almost No Chance	Very Unlikely	Unlikely	Roughly Even Chance	Likely	Very Likely	Almost Certain
1-5%	5-20%	20-45%	45-55%	55-80%	80-95%	95-99%