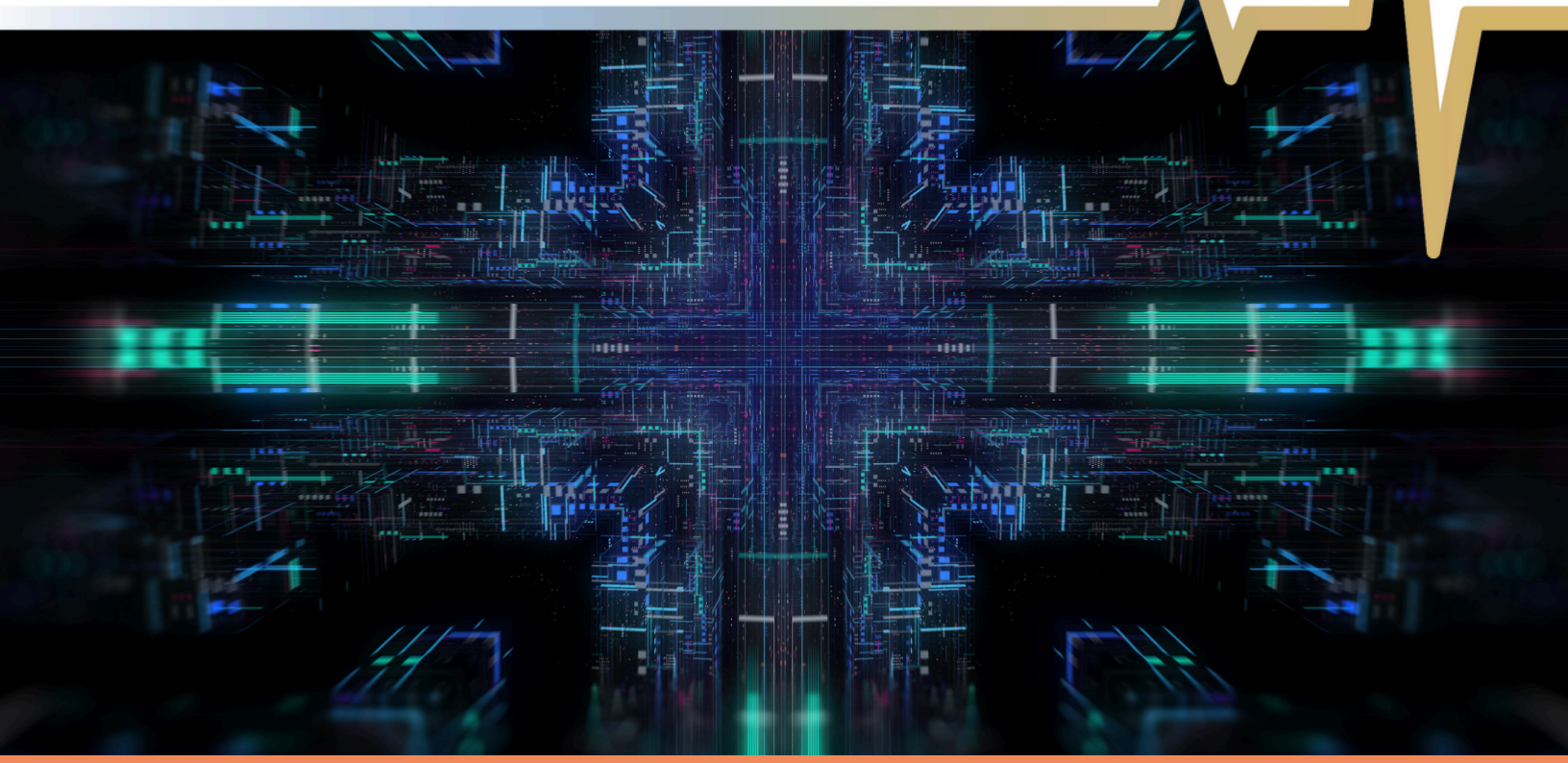
Claude Mythos and its Health Sector Implications

Authors:
Don Methfessel – Sr. IT Security Specialist (Quest Diagnostics)
Sean Hogan – Director of Incident Management and Threat Intelligence (Quest Diagnostics)
Ravi Mani – Chief Information Security Officer (Quest Diagnostics)

Ethan Muntz - Strategic Threat Analyst (Health-ISAC)



Key Judgements

- Claude Mythos Preview possesses unprecedented autonomous offensive capability, including the discovery and exploitation of zero-day vulnerabilities.
- The model's high capability creates a significant risk of misuse, similar to the abuse of tools like Cobalt Strike and Brute Ratel.
- This technology is unlikely to stay in the West. Chinese firms are projected to match Mythos capabilities within 6–12 months.
- Unregulated AI vulnerability discovery models, through unrestricted rollouts or open source offerings, are expected to proliferate the internet by mid-to-late 2026.

Introduction

Claude Mythos is a cutting-edge AI model Anthropic developed as part of its critical Claude Mythos Preview is Anthropic's most capable model to date, described as "a new class of intelligence" built for cybersecurity, autonomous coding, and long-running agentic tasks.[1] Its most significant capability is the autonomous discovery and exploitation of zero-day vulnerabilities across every major operating system and browser. Prior to its restricted release, the model identified thousands of such vulnerabilities, including a 27-year-old flaw in OpenBSD, a 16-year-old single-line bug in the FFmpeg code that had survived five million automated test executions undetected, and a chained Linux kernel privilege escalation exploit, all subsequently patched through coordinated disclosure.[2]

Beyond detection, Mythos Preview can autonomously weaponize these findings into working exploits: it produced 181 working shell exploits against Firefox 147's JavaScript engine in testing, compared to two from its predecessor Opus 4.6, and achieved full control flow hijack on ten separate fully patched targets in Anthropic's internal OSS-Fuzz benchmark.[3]

1. Amazon Web Services (2026, April 7). *Claude Mythos Preview — Amazon Bedrock Model Card*. AWS Documentation. docs.aws.amazon.com/bedrock/latest/userguide/model-card-anthropic-claude-mythos-preview.html
2. Carlini, N. et al. (Anthropic) (2026, April 7). Assessing Claude Mythos Preview's cybersecurity capabilities. Anthropic Red Team Blog. red.anthropic.com/2026/mythos-preview; also Anthropic (2026, April 14), Project Glasswing, anthropic.com/glasswing
3. Carlini, N. et al. (Anthropic) (2026, April 7). Assessing Claude Mythos Preview's cybersecurity capabilities — Firefox 147 benchmark; OSS-Fuzz five-tier benchmark. Anthropic Red Team Blog. red.anthropic.com/2026/mythos-preview

Notably, these capabilities are accessible to non-expert engineers with no formal security training who have used the model to obtain working remote code execution exploits overnight.[4]

Across broader benchmarks, Mythos Preview substantially outperforms Opus 4.6 in every evaluated category. It scores 83.1% on CyberGym (vs. 66.6%), saturates Anthropic's Cybench Capture the Flag (CTF) evaluation at 100%, and achieves 93.9% on Software Engineering Benchmark (SWE-bench) Verified (vs. 80.8%) and 94.6% on Graduate-Level Google-Proof Q&A (GPQA) Diamond (vs. 91.3%).[5] An independent evaluation by the UK AI Security Institute (AISI) found it to be the first model to complete a 32-step corporate network attack simulation end-to-end, a task estimated to require 20 hours of expert human effort, succeeding in 3 of 10 attempts with minimal human direction.[6] AISI qualified its findings, noting that test environments lacked active defenders and that the model's autonomous attack capability applies specifically to "small, weakly defended and vulnerable enterprise systems."[7] The model operates with a 1,000,000-token context window, supports extended reasoning, and is available via the Anthropic API, Amazon Bedrock, Google Cloud Vertex AI, and Microsoft Foundry.[8]

Benchmark	Claude Mythos Preview Score	Opus 4.6 Score
CyberGym	83.1%	66.6%
Cybench Capture the Flag (CTF)	100%	N/A
Software Engineering Benchmark (SWE-bench) Verified	93.9%	80.8%
Graduate-Level Google-Proof Q&A (GPQA) Diamond	94.6%	91.3%

4. Carlini, N. et al. (Anthropic) (2026, April 7). *Assessing Claude Mythos Preview's cybersecurity capabilities — Non-expert accessibility*. Anthropic Red Team Blog. red.anthropic.com/2026/mythos-preview
5. Anthropic (2026, April 7). Claude Mythos Preview System Card. Official System Card, §6 Capabilities. anthropic.com/claude-mythos-preview-system-card Carlini, N. et al. (Anthropic) (2026, April 7). Assessing Claude Mythos Preview's cybersecurity capabilities. Anthropic Red Team Blog. red.anthropic.com/2026/mythos-preview; also Anthropic (2026, April 14), Project Glasswing, anthropic.com/glasswing
6. UK AI Security Institute (AISI) (2026, April 13). Our evaluation of Claude Mythos Preview's cyber capabilities — TLO Cyber Range results. AISI Official Blog. aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities
7. UK AI Security Institute (AISI) (2026, April 13). Our evaluation of Claude Mythos Preview's cyber capabilities — Implications. AISI Official Blog. aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities
8. Amazon Web Services (2026, April 7). Claude Mythos Preview — Amazon Bedrock Model Card, Technical Specifications. AWS Documentation. docs.aws.amazon.com/bedrock/latest/userguide/model-card-anthropic-claude-mythos-preview.html

Risk Assessment

Mythos, like other dual use cybersecurity technologies, is operating under a strict know-your-customer (KYC) model. In fact, the model has only been rolled out to 40 thoroughly vetted organizations through Project Glasswing.

Should the rollout stay within the authorized parties, it would be a success and set the stage for a wider rollout. However, it could create a force multiplier for criminal elements if leaked. Some other tools used for legitimate security purposes that were abused by threat actors include the command-and-control frameworks Cobalt Strike and Brute Ratel. Both tools were created to help enterprise red teams better assess security at their organizations, but both ended up being used by threat actors, by purchasing cracked versions of the software or getting legitimate access by misleading the developers, to launch malicious cyber attacks. Fortra, the company behind Cobalt Strike, teamed up with Health-ISAC and Microsoft to build a legal case to disrupt unauthorized access to Cobalt Strike.[9]

Should the rollout of Claude Mythos follow a similar trajectory, it could jeopardize health sector security like the abuse of Cobalt Strike and Brute Ratel. Concerningly, there are already claims circulating of unauthorized parties gaining access to Mythos. On April 21, Bloomberg reported that a group of users on Discord had gained access to the model through alleged inside access at a third party of Anthropic, locating the preview model using the naming schemes found in the breach data from AI startup Mercor.[10] Anthropic is aware of this claim and is investigating, although preliminary reports suggest the breach is not legitimate.

9. Combating Cybercriminals: Disrupting Abused Security Tools | Security Insider. (2023). Microsoft.com. https://www.microsoft.com/en-us/security/security-insider/risk-management/stopping-cybercriminals-from-abusing-security-tools/
10. Report: Discord Group Uses Claude's Supposedly Secret Mythos. (2026). Govinfosecurity.com. https://www.govinfosecurity.com/report-discord-group-uses-laudes-supposedly-secret-mythos-a-31484

Systemic Implications

While access to Claude Mythos remains restricted to its initial rollout group in Project Glasswing, recent reports suggest that Chinese firms could match its capabilities within 6-12 months. Leading the charge in the Chinese market is 360 Digital Security Group, which utilized its internally developed multi-agent collaborative vulnerability discovery system to win first place at the January 2026 Tianfu Cup, a major Chinese vulnerability identification contest.

The company claims its AI system has evolved into the core engine of its research, identifying nearly 1,000 vulnerabilities, including over 50 high-severity flaws across client-rich products like Microsoft Office, Windows, and OpenClaw. Unlike the autonomous reasoning of Claude Mythos, the 360 system acts as a structured pipeline designed for high-efficiency reproduction and analysis, similar to Google's Big Sleep.[11] Most critically, structural mechanisms in China's legal framework funnel these discoveries directly into government pipelines, making it almost certain that a Chinese vulnerability identification AI model could become operationalized for nation-state offensive cyber operations much faster than Western counterparts.

Adding to the concern of unregulated AI vulnerability discovery models, open-source offerings through Hugging Face and other platforms will likely begin to be offered in mid- to late 2026. These models will have no effective regulation and will make AI vulnerability discovery more accessible to cybercriminals.

11. Benincasa, E. (2026, April 22). Chinese Firm Claims AI-Driven Bug Discovery Near Claude Mythos Scale. Nattothoughts.com; Natto Thoughts. https://www.nattothoughts.com/p/where-is-china-in-ai-driven-vulnerability

Open-source implementations of the theoretical architecture used to build Mythos are already available on open-source software hubs. OpenMythos by kyegomez is an attempted reconstruction of the Recurrent Depth Transformer (RDT) architecture used in Claude Mythos. It is available on GitHub, has 10,000 stars, and has seen significant cybersecurity media adoption, highlighting considerable interest.[12] Compounding this concern, OpenAI's GPT 5.5 has been performing exceptionally well in offensive security benchmarking. Specifically, its capabilities were found to excel in black box testing. ChatGPT 5.5 is available to premium users OpenAI users, making this technology much more accessible.[13]

Additional Thoughts for Infosec Leadership

The rise of AI, and especially agentic AI, changes the operational math of enterprise security. As models like Claude Mythos Preview demonstrate, autonomous systems can accelerate reconnaissance, vulnerability research, exploit construction, and even end-to-end attack execution, compressing the timeline from "finding" to "weaponizing" from weeks into hours. That shift is strategically important: many security programs are still optimized for human-paced adversaries and relatively predictable remediation cycles, but we're moving toward an environment where attackers can industrialize exploitation with less expertise, less time, and greater scale. For security leaders, this means resilience increasingly depends on how quickly the organization can translate threat signals into risk-reducing action, not just how well it can detect and respond.

In practice, patch management becomes a frontline strategy problem, not a back-office hygiene task. Traditional enterprise patching, ticket queues, change windows, dependency mapping, cross-team approvals, and "perfect conditions" governance, was built to maximize stability, not speed, and agentic adversaries will exploit that gap. Infosec and IT operations will need to treat remediation as a core capability: ruthlessly prioritizing based on exploitability and exposure, engineering for rapid and repeatable change (automation, testing-at-scale, and pre-approved emergency paths), and reducing blast radius when patching can't happen immediately through segmentation and other compensating controls. The uncomfortable reality is that defenders won't have the luxury of waiting because adversaries will have increasingly fewer technical barriers to acting immediately.

12. kyegomez. (2025). GitHub - kyegomez/OpenMythos: A theoretical reconstruction of the Claude Mythos architecture, built from first principles using the available research literature. GitHub. Benincasa, E. (2026, April 22). Chinese Firm Claims AI-Driven Bug Discovery Near Claude Mythos Scale. Nattothoughts.com; Natto Thoughts. https://www.nattothoughts.com/p/where-is-china-in-ai-driven-vulnerability
13. XBOW - GPT-5.5: Mythos-Like Hacking, Open To All. (2026). Xbow.com. https://xbow.com/blog/mythos-like-hacking-open-to-all

Conclusion

Claude Mythos Preview represents a genuine inflection point in AI-driven cybersecurity. Its ability to autonomously discover decades-old vulnerabilities, construct working exploits, and complete complex network attack simulations with minimal human direction marks a qualitative leap beyond anything previously demonstrated by a commercial AI model. Deployed responsibly through Project Glasswing's tightly controlled rollout, it has the potential to fundamentally shift the defensive posture of critical infrastructure, finding and closing vulnerabilities faster than any human team could.

The risk, however, is proportional to the capability. The parallels to Cobalt Strike and Brute Ratel are not abstract; they are a documented pattern of legitimate security tools becoming weapons in adversarial hands. With early unauthorized access claims already surfacing, a Chinese competitor potentially already operational, and open-source reconstructions gaining traction on public platforms, the window in which Mythos-level capability remains contained to vetted defenders may be narrower than its architects intend. Whether the technology ultimately strengthens global security or accelerates the threats it was built to prevent will depend entirely on how effectively that window is managed.