JOINT INFORMATION AWARENESS BULLETIN
ANTHROPIC CLAUDE MYTHOS

28 April 2026

Executive Summary

On April 14, 2026, Anthropic, the company behind the Claude Artificial Intelligence (AI) models, released a preview of its most advanced frontier model, Claude Mythos, to select partners through Project Glasswing. Anthropic indicates they are not currently releasing Claude Mythos publicly as its cyber capabilities may have the ability to create misuse risk. Anthropic reports that Claude Mythos demonstrates a major jump in cyber capability, including the ability to find unknown software vulnerabilities, develop exploit methods, and carry out multi-step cyber tasks under controlled conditions.

Open-source reporting indicates three individuals gained unauthorized access to Claude Mythos. However, Anthropic reports that the aforementioned access has been revoked, and reports no other access control concerns.

This bulletin outlines the cyber risks associated with Claude Mythos, describes how a model with these capabilities may be misused by Cyber Threat Actors (CTAs), addresses how California public-sector entities and partners could use similar capabilities to improve cybersecurity posture, and provides practical recommendations to improve readiness.

What is Claude Mythos

Anthropic describes Claude Mythos as its most capable frontier model to date. Anthropic restricts access to Claude Mythos because the model's cyber capabilities could create misuse risk. Anthropic presents Claude Mythos as having an unusually strong ability to identify software vulnerabilities, develop exploit methods, and carry out complex multi-step cyber tasks under controlled conditions.

According to Palo Alto Networks (PAN), Anthropic describes Claude Mythos as an in-development model that could pose substantial cybersecurity risk if misused. In that context, Claude Mythos should be understood not as a public consumer product, but as a restricted advanced model associated with high-end cybersecurity capability and limited partner access through Project Glasswing.

Cal OES
GOVERNOR'S OFFICE OF EMERGENCY SERVICES

JOINT INFORMATION AWARENESS BULLETIN
ANTHROPIC CLAUDE MYTHOS

28 April 2026

What is Project Glasswing

Project Glasswing is Anthropic's early-access initiative that manages the risk of misuse by restricting access to secure partner and vendor environments with strong access controls. This enables select partner organizations to use Claude Mythos in controlled defensive cybersecurity work.

On April 22, 2026, Reuters reported that unauthorized users accessed Claude Mythos through a third-party vendor environment rather than through a broad public leak. Anthropic reports that the unauthorized access has been revoked and reports no other access control concerns.

Assessed Risks of Claude Mythos

Anthropic presents Claude Mythos as a frontier model with both defensive and offensive cybersecurity relevance. It can help identify and support remediation of software vulnerabilities, but those same capabilities could also help develop methods to exploit those vulnerabilities. Claude Mythos demonstrates a major jump in cyber capability relative to prior models, including the ability to autonomously discover and exploit zero-day vulnerabilities in major operating systems and web browsers.

The most immediate risk is **autonomous vulnerability discovery**, meaning the model can help identify software vulnerabilities with limited human direction. In authorized testing, Anthropic reports that Claude Mythos could identify vulnerabilities in open and closed-source software and, in many cases, develop those findings into exploits.

For California public-sector leaders, the practical implication is compressed decision time; agencies may have less time to identify exposed systems, prioritize critical patches, coordinate with vendors, and protect vulnerable systems from becoming operational incidents. Anthropic reports that Claude Mythos performed at or near the upper limit of nearly all existing cyber capability evaluations, suggesting those tests may no longer fully measure its capability and indicates a major jump in how quickly software vulnerabilities can be identified and acted on.

JOINT INFORMATION AWARENESS BULLETIN
ANTHROPIC CLAUDE MYTHOS

28 April 2026

Assessed Risks of Claude Mythos Continued

A second risk is exploit weaponization, or the process of turning identified software vulnerabilities into usable attack methods that can be deployed against specific targets. Anthropic claims Claude Mythos's capabilities are not limited to identifying vulnerabilities. It can help turn technical findings into working exploits and solve realistic cyber tasks that require linking multiple vulnerabilities together. Mythos-like capability could reduce the time and expertise needed to move from vulnerability discovery to usable attack capability. For leaders, the practical implication is that gaps in asset visibility, patching, vendor oversight, and incident response could become operational, financial, and reputational risks faster than before.

Claude Mythos creates broader pressure on public-sector defense as AI-enabled vulnerability discovery and exploit development may reduce the time defenders have to patch, detect, and respond. California Department of Technology (CDT) guidance frames rapid advances in AI-enabled vulnerability discovery, exploit generation, and autonomous attack orchestration as a structural shift in the cyber threat environment. Public sector entities should consider a structural shift to their cyber threat environments to respond to rapid AI advances. This is challenging to accomplish given the pace at which resources are available to public-sector entities, and maintain long patch cycles as downtime for their systems can carry significant public consequence.

As highlighted by Mandiant, the existence of a model with this level of capability poses a security risk to all accessible systems, including developer environments, orchestration frameworks, and exposed application programming interface (API) keys. For California entities, this broadens the risk picture from model capability alone to the security of the software, access points, and supporting infrastructure connected to advanced AI use.

The April 22, 2026, unauthorized-access reporting does not change the core assessment of Claude Mythos's cyber capability. It does, however, reinforce that preview environments, vendor portals, and related access-control processes are part of the risk picture when highly capable models are shared with outside partners.

JOINT INFORMATION AWARENESS BULLETIN
ANTHROPIC CLAUDE MYTHOS

28 April 2026

Potential Utilization of AI For Cyber Network Defense

As Cyber Threat Actors (CTAs) increasingly use advanced AI models, like Claude Mythos, to scale cyberattacks, California public sector entities and partners face a clear need to respond with comparable and adaptable defensive capabilities. With the right controls in place, agentic AI, meaning AI that can carry out multi-step tasks with some autonomy, utilizing advanced large language models (LLMs) like Claude Mythos, can help identify and prioritize vulnerabilities faster, reducing response timelines from days toward hours.

For secure software development, AI-enabled tools and enterprise security platforms, such as GitHub Advanced Security or GitLab Ultimate, can help identify insecure code earlier, improving software quality before it reaches production. Advanced models like Claude Mythos can be used to conduct code review during software development cycles. Similar AI-enabled tools can also help interpret emerging threat intelligence and guide remediation efforts, supporting faster and more effective fixes without introducing new risks.

Beyond vulnerability management, AI systems like Claude Mythos can continuously analyze security data and historical activity to support more effective threat hunting and help uncover intrusions that may have bypassed traditional defenses. Behavior-based detection, a strength of AI tools, supports that effort by identifying suspicious activity based on what it does, rather than only matching it against known threat signatures. This can make it easier to detect new or AI-driven attack methods.

For California public sector entities, the priority should be:

- Applying AI-enabled defensive tools to improve vulnerability response speed, visibility into suspicious activity, and resilience of critical systems; and
- Aligning with policy and security requirements consistent with existing guidance and the Cal-Secure Roadmap from CDT which emphasizes faster and adaptable patching, stronger visibility, and better coordination.

Overall, these capabilities should support faster defensive decision-making while remaining grounded in governance, trained personnel, and established security controls.

JOINT INFORMATION AWARENESS BULLETIN
ANTHROPIC CLAUDE MYTHOS

28 April 2026

Recommendations

Immediate Actions

- Enforce basic cyber hygiene (e.g. Multi-Factor Authentication, change default passwords)
- Secure privileged accounts and remote access pathways
- Reduce unnecessary public exposure of administrative interfaces
- Confirm clear incident reporting and escalation procedures
- Ensure relevant personnel are enrolled in Cal-CSIC alerting and information-sharing channels

Next 30 to 90 Days

- Reduce patching timelines for critical vulnerabilities
- Strengthen network segmentation to limit lateral movement
- Examine managed service provider and software-as-a-service dependencies
- Improve asset inventory and Software Bills of Materials

Longer Term

- Strengthen governance and risk tracking
- Improve detection engineering and response workflows
- Build or source vulnerability operations capability
- Strengthen workforce resilience
- Build resilience for under-resourced organizations

Sources of Cited Information

Anthropic Claude Mythos Preview– Official Claude Mythos capability and risk overview

Mandiant Threat Intelligence – Cyber threat reporting and trends

Palo Alto Networks (PAN) – Claude Mythos incident reporting and analysis

Reuters – Claude Mythos Access Reporting – Claude Mythos incident reporting

Resources

FBI-IC3 – Report a cyber crime

State Threat Assessment System – Homeland Security Information and Intelligence Sharing

Cal-CSIC Transparency in Frontier AI Act (TFAIA) Reporting – Submit frontier AI safety incident data

JOINT INFORMATION AWARENESS BULLETIN
ANTHROPIC CLAUDE MYTHOS

28 April 2026

Sharing, Reporting, and Support

Sharing information and intelligence to state, local, federal, and private partners helps organizations across California stay better informed and defended.

For questions, concerns, support, or interest in sharing information or intelligence in any cyber related matter, contact the Cal-CSIC via:

916.636.2997

calcsic@caloes.ca.gov

Report suspicious or suspected foreign activity and incidents to:

- Federal Officials at CISA and/or the FBI as applicable
- Coordinate with your regional Fusion Center
- SHARE Information, intelligence, and IOCs with ISACs