INTERNATIONAL JOURNAL OF ENGINEERING MANAGEMENT SCIENCE
journal homepage: www.ijems.online/index.php/ijems/index

The Emergence of Anthropic's Claude Mythos Preview: Zero-Day Vulnerabilities, Defensive AI Initiatives, and Persistent Cybersecurity Risks

Dr. Ramesh.S[1], Akriti Yadav[2], Md Shaan Nishat[3], Trisha Roy Choudhury[4], Srilakshmi J S[5]

[1]Professor, Indus Business Academy Bangalore.

[2]PGDM 2024-26, Indus Business Academy Bangalore.

[3]PGDM 2025-27, Indus Business Academy Bangalore.

[4]PGDM 2025-27, Indus Business Academy Bangalore.

[5]PGDM 2025-27, Indus Business Academy Bangalore.

Keyword

Claude Mythos,
Zero-Day Vulnerabilities,
Anthropic AI,
Frontier AI Risks,
Cybersecurity Threats,
Persistent Zero-Days,
India Banking Security,
AI Philosophical Emergence,
Defensive AI Initiatives,
Cybercrime Democratization,
Critical Infrastructure Protection,
RBI Cybersecurity Response.

Abstract

The April 2026 release of Anthropic's Claude Mythos Preview marked a pivotal moment in cybersecurity. This frontier AI model autonomously identified thousands of previously unknown zero-day vulnerabilities across major operating systems, web browsers, and foundational software libraries. In response, Anthropic launched Project Glasswing, a restricted-access defensive program partnering with global tech and finance leaders. This paper synthesizes the technical, philosophical, and geopolitical implications of Mythos, with reference to the high-level emergency meeting convened by India's Finance Minister Nirmala Sitharaman with banks on 23 April 2026. It explores why nation-states and sophisticated hackers had not previously uncovered these flaws at scale, examines hypothetical misuse scenarios, and highlights a critical limitation: even Mythos cannot identify all zero-days, leaving residual risks in an accelerating AI-driven arms race.

1. INTRODUCTION

Global cybersecurity concerns have intensified following the emergence of Anthropic's advanced AI model called Claude Mythos (often shortened to Mythos). The "meeting of FM with banks" was a high-level emergency discussion held by India's Finance Minister Nirmala Sitharaman on April 23, 2026.

Anthropic (the company behind the Claude AI family) developed Mythos as a highly advanced AI with exceptional capabilities in cybersecurity. In testing, it reportedly identified thousands of major zero-day vulnerabilities across every major operating system, web browser, and software stack. While Anthropic built it mAInly for defensive purposes (helping companies fix flaws before attackers exploit them) and has not released it publicly (access is tightly restricted to select partners under "Project Glasswing"), experts and regulators worry about its potential misuse.

Mythos can map, analyze, and weaponize software weaknesses far faster than human hackers. This could enable sophisticated, large-scale cyberattacks on critical systems like banking software, payment networks (UPI, etc.), customer data repositories, and legacy infrastructure common in banks. The fear is an "unprecedented" rise in cybercrime, fraud, data breaches, and systemic financial instability. Finance ministers, central bankers, and regulators worldwide (including at recent IMF meetings) have described it as a potential game-changer for cyber threats — comparable to an "unknown unknown" risk.

In response, Finance Minister Nirmala Sitharaman chAIred a high-level meeting in New Delhi with CEOs/heads of major public and private sector banks, Reserve Bank of India (RBI), Ministry of Electronics & IT (MeitY) officials, and IT Minister Ashwini VAIshnaw. Key points discussed included pre-emptive strengthening of cybersecurity across banking systems and customer data protection, development of a coordinated institutional mechanism (via Indian Banks' Association – IBA) for rapid response and real-time threat intelligence sharing, immediate engagement of top cybersecurity experts, and prompt reporting of suspicious activity. The meeting viewed Mythos as both a risk and a potential opportunity for the fintech ecosystem (e.g., using AI defensively). The meeting was prompted by global alerts after Anthropic's announcements and similar discussions in other countries (e.g., Japan's Finance Minister also met banks to address the Mythos threat). India's government and RBI are also in touch with global regulators and have reportedly sought clarifications from Anthropic. Banks have been told to assess their exposure and ensure no impact on customer deposits or services.

For customers and banks in India, the bottom line is reassuring: bank accounts are not under any immediate known attack — this is proactive risk management. Banks are expected to accelerate upgrades in firewalls, AI-based threat detection, patching, and monitoring in the coming weeks and months. The situation continues to be monitored closely by the RBI and the Finance Ministry.

This paper builds on these developments. Its central philosophical idea is that the "something" that prevented nation-states and hackers from discovering these zero-days earlier was not conspiracy, but incentive structures, bounded rationality, and the inherent complexity of software systems — until frontier AI removed the veil.

2. BACKGROUND

Zero-Days and the Mythos Breakthrough A zero-day vulnerability is a flaw in software, hardware, or firmware unknown to the vendor and unpatched at the time of discovery or exploitation. Historically, such flaws have persisted for years — even decades — in widely audited codebases. Anthropic's Claude Mythos Preview was not explicitly trAIned for offensive cybersecurity. Its zero-day capabilities emerged as a side-effect of massive

scaling in general reasoning, coding, and autonomous simulation. The model autonomously mapped execution paths, identified subtle logic errors, and chAIned vulnerabilities across massive codebases. Anthropic chose not to release it publicly due to catastrophic misuse potential and instead restricted access while publicly disclosing the risk.

3. PROJECT GLASSWING

A Defensive-First Response Recognizing the dual-use dilemma, Anthropic launched Project Glasswing. This initiative grants controlled, early access to Mythos Preview to a vetted consortium of major tech and finance organizations. The explicit goal is to enable partners to scan and remediate foundational systems before malicious actors can weaponize equivalent capabilities. In India, this directly informed the RBI and MeitY's guidance to banks.

4. PHILOSOPHICAL AND STRUCTURAL REASONS FOR PRIOR NON-DISCOVERY

Why did nation-states with multi-billion-dollar cyber budgets and well-resourced criminal groups not systematically uncover these thousands of zero-days earlier? The answer lies in incentive misalignment, bounded rationality, and the nature of complexity — not in any deliberate suppression. Nation-states prioritize targeted offensive capabilities and hoard high-value zero-days rather than exhaustively audit global software. Criminal hackers optimize for quick, low-effort returns rather than investing in frontier general AI research.
Mythos succeeded because it emerged from broad, curiosity-driven scaling in a private lab whose incentives aligned with broad testing and partial disclosure. This reveals a deeper philosophical point: powerful instruments do not merely solve problems — they expose the hidden structure of reality itself.

5. IMPLICATIONS FOR CRITICAL INFRASTRUCTURE:

The Indian Banking Context India's 23 April 2026 meeting exemplified proactive risk management. Banks were urged to strengthen firewalls, deploy AI-based threat detection, patch legacy systems, and report suspicious activity immediately. The discussion framed Mythos as both a threat and an opportunity for the fintech ecosystem.

6. HYPOTHETICAL RISKS OF MISUSE

If Anthropic had pursued secret offensive use (analogous to Sam Bankman-Fried's handling of FTX customer funds), or if core researchers had defected to major cybercrime syndicates, the outcome would have been catastrophic. A Mythos-equivalent in adversarial hands could automate large-scale zero-day weaponization, targeting banks, payment networks, and critical infrastructure. These scenarios highlight the asymmetry: offense scales faster than defense once frontier AI is involved.

7. Limitations of Mythos

Unidentified Zero-Days RemAIn Despite its breakthrough, Mythos is not omniscient. Several categories of zero-days may still lie beyond its current reach, including extremely obscure or AIr-gapped systems, emergent flaws in newly built security layers, hardware-level vulnerabilities, and complex multi-system interactions.
In short, Mythos reveals what was latent in known codebases; it does not guarantee completeness across the entire attack surface. Future models will likely expose the next layer, perpetuating the arms race.

8. CONCLUSION

Anthropic's Claude Mythos Preview and Project Glasswing represent both a warning and a rare defensive window. The 23 April 2026 Indian banking meeting underscores the urgency for coordinated responses. Philosophically, the episode reveals how ordinary constrAInts of human and institutional priorities kept systemic weaknesses hidden for decades. Mythos removed one veil; future advances will remove others. Policymakers, banks, and institutions must treat this not as a one-time patch cycle but as the beginning of an AI-accelerated cybersecurity era. Preparedness, not panic, is the rational response.

9. AUTHOR(S) CONTRIBUTION

The writers affirm that they have no connections to, or engagement with, any group or body. That provides financial or non-financial assistance for the topics or resources covered in this Manuscript.

10. CONFLICTS OF INTEREST

The authors declared no potential conflicts of interest with respect to the research, authorship, And / or publication of this article.

11. PLAGIARISM POLICY

All authors declare that any kind of violation of plagiarism, copyright and ethical matters will\ Take care by all authors. Journal and editors are not liable for aforesAId matters.

12. SOURCES OF FUNDING

The authors received no financial AId to support for the research.

References:

[1] Anthropic. (2026). Claude Mythos Preview Technical Report.
[2] Government of India, Ministry of Finance. (2026). Press briefing on high-level meeting with banks, 23 April.
[3] Project Glasswing Consortium materials (2026).