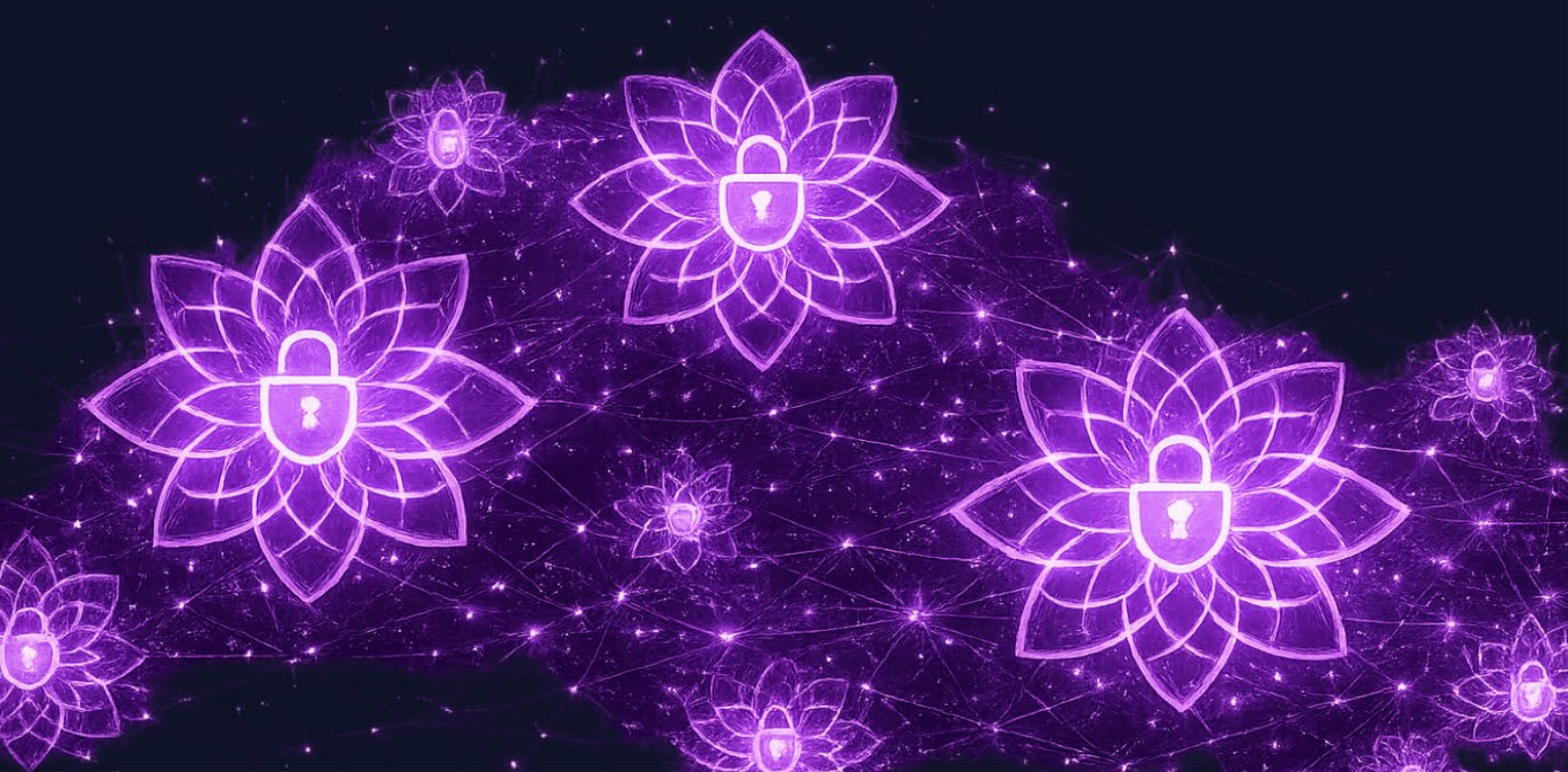
FIOR

Authentication for the Intelligent Age

Fior defence against

Claude Mythos Autonomous AI Exploit Capability

Prepared for:	Technical Readers
Author:	FIOR Group Limited
Classification:	For Publication
Date published:	April 2026



Executive Summary

On April 7, 2026, Anthropic announced Claude Mythos Preview, a frontier AI model with unprecedented cybersecurity capabilities. In internal testing, Mythos Preview autonomously discovered and exploited zero-day vulnerabilities across every major operating system and web browser, including a 27-year-old bug in OpenBSD — an OS renowned for its security.

This represents a step-change in the threat landscape. AI agents can now perform vulnerability research and exploit development that previously required expert human professionals. Non-security engineers using Mythos obtained working remote code execution exploits overnight.

The Threat: What Mythos Can Do

> **"Engineers at Anthropic with no formal security training asked Mythos Preview to find remote code execution vulnerabilities overnight, and woke up the following morning to a complete, working exploit."**

Autonomous Zero-Day Discovery

Mythos Preview found previously unknown vulnerabilities in:

- Every major operating system (Linux, Windows, macOS, FreeBSD, OpenBSD)
- Every major web browser (Chrome, Firefox, Safari, Edge)
- Core system libraries and cryptography implementations
- Virtual machine monitors (hypervisor escape)
- Smart phone lock screens

Autonomous Exploit Development

Beyond finding bugs, Mythos autonomously constructs sophisticated exploit chains:

- JIT heap sprays chained with sandbox escapes in web browsers
- 20-gadget ROP chains split across multiple network packets for remote root access
- Kernel privilege escalation via subtle race conditions and KASLR bypasses
- Multi-vulnerability chains requiring no human intervention

Scale of Improvement

Metric	Opus 4.6 (Previous)	Mythos Preview
Firefox exploit success	Near 0% (2 of hundreds)	181 of hundreds (+ 29 partial)
OSS-Fuzz Tier 5 (full hijack)	0	10 separate targets
Zero-day discovery	Limited	Across all major platforms
Non-expert usability	Minimal	Working RCEs overnight

How FIOR Defends Against Mythos-Class Threats

FIOR's pre-authorization model is uniquely positioned against autonomous exploit agents because it controls what agents *can do* regardless of what they *know how to do*. Even if an agent discovers a zero-day, FIOR's cryptographic constraints prevent exploitation.

Mythos Attack Vector	FIOR Defence Layer	Enforcement	Result
Agent probes internal systems	Domain Allowlist	Certificate constrains endpoints	Blocked
Exploitation on unauthorized target	Operation Block-list	shell_exec, debug_attach blocked	Blocked
Compromised agent pivots laterally	Certificate Identity Binding	Unique scoped identity per agent	Contained
Agent exfiltrates exploit payloads	Rate Limiting + Anomaly	Unusual traffic patterns flagged	Detected
Agent runs overnight unsupervised	Human-in-the-Loop	High-risk actions require approval	Held
Exploit chain spans services	Per-Action Crypto Signing	No replay, no forging	Prevented
Zero-day on protected service	Sub-Second Revocation	Global agent termination	Terminated
Reconnaissance scanning	Behavioral Velocity Limits	Systematic probing detected	Quarantined

The Core Principle: It is easier to prevent unauthorized actions than to detect malicious intent. The certificate is the control boundary — not the prompt, not the model, not the agent's "intent."

Defence in Depth: The Four Gates

- Gate 1 — Pre-Issuance Vetting: Operators are verified before receiving certificates. Mythos-powered agents from unvetted operators never receive credentials.
- Gate 2 — Certificate Constraints: Allowed domains, blocked operations, rate limits, and human approval requirements are cryptographically embedded. No override possible.
- Gate 3 — Runtime Enforcement: Every request is validated against enterprise policies. Exploitation attempts are blocked in real-time.
- Gate 4 — Post-Hoc Revocation: If anomalous behavior is detected, sub-second global revocation terminates all active sessions across every gateway.

Recommended Enhancements for Mythos-Era Security

Exploit Pattern Detection [HIGH PRIORITY]

Certificate constraints that specifically block vulnerability-research behaviors: memory probing patterns, fuzzing-like request sequences, debugger attachment, and unusual syscall patterns.

Behavioral Velocity Limits [HIGH PRIORITY]

Beyond rate limiting — detect when an agent systematically scans endpoints or tests input boundaries. Classic pre-exploitation reconnaissance triggers automatic quarantine.

Sandbox Attestation [MEDIUM PRIORITY]

Require agents to attest they are running inside approved execution environments with verified configurations.

Granular Capability Narrowing [MEDIUM PRIORITY]

Post-Mythos certificates should specify allowed API endpoints, methods, and payload schemas.

Cross-Gateway Threat Intelligence [HIGH PRIORITY]

When one gateway detects Mythos-style probing, broadcast the pattern signature to all gateways. Collective immune response.

Quantum-Safe Hardening [MEDIUM PRIORITY]

Ensure all certificate signatures use CRYSTALS-Dilithium or FALCON to prevent future quantum attacks from forging agent identities.

Industry Context: Project Glasswing

Anthropic has launched Project Glasswing with AWS, Apple, Google, and Microsoft to accelerate vulnerability discovery and patching. While complementary, Glasswing addresses the **supply side** (finding and fixing bugs). FIOR addresses the **demand side** (preventing exploitation during the patch gap).

Only FIOR protects during the window between vulnerability discovery and patch deployment.