CyRisk

INTELLIGENCE BRIEF

CYRISK INTELLIGENCE BRIEF · APRIL 2026

THE GLASSWING ERA: AI-POWERED VULNERABILITY DISCOVERY AND WHAT IT MEANS FOR CYBER UNDERWRITING

Anthropic's Mythos Preview and Project Glasswing have redrawn the cyber threat landscape. This brief explains what changed, what it means for risk selection and pricing, and what the insurance market can do to respond.

PUBLISHED
April 2026

AUDIENCE
Cyber Underwriters · Portfolio Risk Managers · Reinsurers

CLASSIFICATION
Public Distribution

INSIDE THIS BRIEF

A concise, human-written briefing on what Anthropic's Claude Mythos Preview and Project Glasswing mean for cyber underwriting, and what the insurance market can do to respond before the next renewal cycle.

EXECUTIVE SUMMARY

THE HARD MARKET ARRIVES BY CHRISTMAS

Since last winter we have been forecasting an 18 to 24-month return to a hard cyber insurance market. Our forecasts are already obsolete. Anthropic's April 7 announcements compressed that timeline from years to months. Carriers, MGAs, and reinsurers should expect a hard cyber market in time for Christmas, and should be repositioning their books now, not after the first wave of post-Glasswing losses arrives.

On April 7, 2026, Anthropic announced Claude Mythos Preview alongside Project Glasswing, a coordinated, industry-wide response to what it described as a watershed moment in cybersecurity. Anthropic's newest model has demonstrated that it can autonomously discover and exploit zero-day vulnerabilities across every major operating system, browser, and critical software stack, at a cost and speed that represent a radical change to what we used to call the "ever-changing cyber threat landscape." (Ah, the good old days.)

Many believed the announcement was simply a marketing tactic. That interpretation completely misses the importance of these events. On the *same day*, Treasury Secretary Scott Bessent and Federal Reserve Chair Jerome Powell convened an unannounced emergency meeting at Treasury headquarters with the CEOs of the country's most systemically important financial institutions: Citigroup, Morgan Stanley, Bank of America, Wells Fargo, and Goldman Sachs. The subject was the cyber risk posed by Mythos. Marketing ploys do not get that kind of response.

The April 7 emergency meeting at Treasury was not an isolated event. The International Monetary Fund, the European Central Bank, and other regulators and banking institutions around the world have been meeting to assess the systemic risk posed by Mythos-class LLMs to the global financial system. They are all taking these events deadly serious, and rightly so.

For cyber underwriters and their enterprise clients, this is not background noise. The attack surface has expanded materially, the cyber kill chain has collapsed, the economics of exploitation have shifted dramatically, and the signals that define a well-managed vs. poorly-managed risk profile need to be updated accordingly. All signs point to a heavy increase in the frequency and severity of cyber claims.

THE AUTONOMOUS PARADOX

Anthropic's Mythos Preview is an accidental apex cyber predator.

9,000%
Increase in working Firefox exploits Mythos produced vs. the prior-generation model.

<$2K
API cost to autonomously develop a working N-day exploit in hours, not weeks.

27 yrs
Age of the OpenBSD vulnerability Mythos found without human guidance.

72%
Autonomous exploit-development success rate on the CyberGym benchmark.

01 THE GLASSWING ANNOUNCEMENTS

Anthropic released two documents on April 7, 2026: a technical deep-dive on Mythos Preview's offensive cybersecurity capabilities, and the Project Glasswing announcement: a 12-partner industry coalition (now expanded to over 50 partners) including AWS, Apple, Microsoft, Google, CrowdStrike, Palo Alto Networks, Cisco, JPMorgan Chase, Broadcom, NVIDIA, and the Linux Foundation.

It is important to note that Anthropic didn't set out to build Mythos as a digital weapon. They didn't train it with the explicit goal of creating a super-hacker. But teaching it to be a master software engineer made it exceptionally good at finding the exact points where software security breaks down. AI security researchers call this the autonomous paradox: the new offensive capability was not programmed; it *emerged*. When Anthropic realized they had created an apex predator on the cyber battlefield, they didn't just release it. They locked it up and formed Project Glasswing.

> *"Teaching a model to be a master software engineer made it exceptionally good at finding the exact points where software security breaks down."*

KEY FINDINGS FROM THE MYTHOS PREVIEW TECHNICAL REPORT

- **Zero-days, autonomously.** Critical vulnerabilities found in every major OS (Windows, Linux, MacOS, FreeBSD, OpenBSD) and every major browser.
- **27-year-old OpenBSD bug.** A remote machine-crash vulnerability was found without human guidance.
- **Firefox 147 JavaScript engine.** Claude Opus 4.6 produced 2 working exploits. Mythos Preview produced **181**, a **9,000% increase** in offensive capability in a single generation.
- **16-year-old bug in a ubiquitous video software suite** missed by 5 million automated test runs was caught by Mythos in hours.
- **Linux kernel exploit chains** producing full root access were developed end-to-end by the model.
- **N-day development time collapsed** from weeks (expert researcher) to hours, at < $2,000 API cost.
- **Benchmarks:** 83.1% on CyberGym vs. 66.6% for the prior best model; 72.4% autonomous exploit development; 92.1% TerminalBench 2.0 with adaptive thinking.

Source: *Mythos Preview Technical Report* (Anthropic, April 7, 2026) and the *Project Glasswing Coalition Announcement*. Figures as reported by Anthropic; interpretation by CyRisk.

02 FOUR BIG SHIFTS

01 The N-Day Window Has Collapsed

Historically, the lag between a vulnerability being disclosed and a working exploit being weaponized gave enterprise security teams *weeks* to patch. A 30-day patching cycle was acceptable, typically enough to prevent an attack. That window has pretty much gone the way of the Dodo bird. Prior AI models have already proven to be extremely effective tools to help find security vulnerabilities. Anthropic's Mythos Preview demonstrated the ability to produce a functional exploit autonomously, in hours, for less than $2,000. It is in a different class altogether.

Every cyber loss model that relies on the distinction between a disclosed-but-unpatched vulnerability and an actively exploited vulnerability must be recalibrated. Patch velocity, how quickly an organization applies fixes after disclosure, is now a critical control to monitor, not a secondary hygiene factor.

IMPLICATION

Organizations with slow patch cycles (30+ day windows, quarterly maintenance schedules, manual update processes) carry materially higher expected loss frequency in a Mythos-era threat environment. This should be reflected in risk selection, attachment points, and pricing.

02 The Expert-Attacker Ceiling Dropped a Lot

Prior to Mythos-class models, creating exploits and chaining multiple zero-days to achieve full system compromise required elite offensive security researchers. This was a small, expensive talent pool historically associated with nation-state actors. Mythos Preview performed this autonomously, without human guidance, across multiple operating systems.

As we all worry about which of our friends' and colleagues' jobs will be supplanted by AI next, "elite black-hat hacker" might not have been at the top of the list. Anthropic noted that non-experts (engineers with no formal security training) were able to get the model to find Remote Code Execution (RCE) vulnerabilities overnight and wake up to working exploits. *The threat actor profile for sophisticated attacks has just expanded, turning script kiddies into potential elite black-hat hackers.*

With Mythos, finding vulnerabilities is easy. In the world of enterprise IT, fixing them is not. What made the benchmark results alarming wasn't just the *volume* of exploits generated. It was *how*. Mythos chained lower-severity flaws, individually classified as minor and deprioritized, into a path to total system takeover. Patching only the most critical bugs is no longer a winning strategy as the National Vulnerability Database (NVD) and the Common Vulnerability Scoring System (CVSS) collapse under the weight of AI-driven discovery. The cyber defender's job just got a lot harder. (See CyRisk's Glasswing Risk Index for a different approach to this challenging situation.)

> *"Underwriters who continue to accede to brokers' threats that they will 'take the business to someone else who doesn't ask so many questions' will be able to warm their hands by the fire of their burning book next winter."*

03 The Open-Source Supply Chain Is More Exposed Than Ever

FFmpeg (used in virtually every major platform for video processing), OpenBSD (common in firewalls and critical network infrastructure), and the Linux kernel underpin an enormous portion of enterprise software. Mythos Preview found critical vulnerabilities in all three: bugs that had remained undiscovered through many years of human review and automated fuzzing.

Jim Zemlin of the Linux Foundation, a Project Glasswing partner, noted that open-source maintainers have historically been left to figure out security on their own, even though open source constitutes the vast majority of code in modern systems. In this new Glasswing Era, enterprise dependency on unaudited or under-resourced open-source components has become a quantifiable, urgent risk factor.

The Glasswing news is not all bad from the cyber defenders' standpoint. Undoubtedly, new code written with the help of a superhero-class AI cybersecurity expert should yield far fewer exploitable bugs. And pointing Mythos at existing code has already proven highly effective at unearthing bugs for defenders to fix, as Bobby Holley, CTO at Firefox recently stated: "This week's release of Firefox 150 includes fixes for 271 vulnerabilities identified during [our] initial evaluation [with Claude Mythos Preview]. For a hardened target, just one such bug would have been red-alert in 2025, and so many at once makes you stop to wonder whether it's even possible to keep up."

But anyone who thinks the net score after all this is attackers 1, defenders 1 has never spent a day toiling in the shoes of an IT worker. As one Redditor 'Hirokage' put it "In our industry, our software stack is full of legacy or small team created apps that most likely will not be updated but be at risk for zero-days. And we have acquired I don't know.. 10 + companies in the last two years, and they also have tons of legacy stuff. ...in my experience it is going to take a LONG time before much of the software out there is patched. Banks or large companies that only use big player software that will be patched? Ok.. that's a small subset of reality." Perhaps Hirokage has not worked IT at a bank or large company. As members of multiple industry cybersecurity and AI working groups, it is our experience that the vast majority of organizations today are still trying to grapple with fundamental AI implementation questions and are not in a position to fix all the bugs, even if patches were made available immediately. Thinking AI can patch everything is also not realistic. There are many reasons for this, but as a simple mental image, consider what would happen if the U.S. decided to fix every pothole on every road all at once. And that's only the computers. Now throw in the firmware embedded in network devices, medical devices, IoT and the list goes on. And let's expand this picture to every country in the world. Maybe that smart refrigerator wasn't such a great idea after all...

This exposure compounds with a newer, less-discussed vector: AI-hallucinated dependency packages. As AI coding assistants become essential tooling, they consistently suggest package names that do not exist. Enterprising attackers register those names with malicious payloads, just like typo-squatters do with expired domains. Researchers have identified over ten thousand AI-hallucinated package names that represent ready-made supply-chain backdoors. Most organizations whose developers use AI coding tools have not deployed any systematic way to screen for them.

RECOMMENDATION

Open-source exposures and AI-generated dependency risks are now both quantifiable underwriting signals. Software Bill of Materials (SBOM) analysis and developer tooling governance should move from "best practice" to required inputs for risk selection and pricing.

04 AI Systems Are Attack Surfaces Too

As enterprises are driven by boards and C-suites to rapidly transition to AI-assisted operations across security, underwriting, finance, customer service, and nearly every modern business function, AI applications, models, and their supporting infrastructure are quickly becoming *primary targets*. Three common attack mechanisms are active today:

A Prompt Injection

Prompt injection attacks increased 540% in 2025, making them the fastest-growing AI threat vector. Unlike traditional systems which separate data from code, LLMs process system instructions and user input through the same natural language interface. An attacker can embed a hidden instruction in a document or email that overrides the model's original instructions entirely. For example, a customer-service AI told to summarize a complaint can be redirected by text hidden in the email body, instructing it to forward the last ten messages to an external server. No code is exploited; the AI follows the new instruction. Indirect prompt injection is a variant where hidden instructions on seemingly innocent websites are ingested by AI scrapers, causing models to ignore guard rails and do things they were not intended to do.

B Data Poisoning

Data poisoning compromises the foundational layer before deployment. If an adversary manipulates training data, they can bake blind spots directly into the model's logic. Google DeepMind researchers demonstrated this on ImageNet (altering pixels invisibly to humans), training a vision system that confidently misidentified dogs as cats. Apply the same technique to an enterprise document repository or a security AI trained on corrupted network logs, and the model learns to ignore specific attack signatures. It looks right at the intrusion and waves it through.

C Agentic Compromise

The explosive popularity of the open source agentic framework, OpenClaw, created a systemic security crisis with the rapid deployment of high-privilege agents with virtually no security, often bypassing corporate IT oversight. The ultra-low-friction adoption, combined with the cyber security illiteracy of the ClawHub marketplace, allowed attackers to flood the ecosystem with thousands of malicious skills that act as pre-authenticated backdoors. Because these AI agents are granted deep access to local terminals, private files, and internal networks, a single compromised skill such as the "What Would Elon Do?" Skill can pivot from a simple chat interface to exfiltrating sensitive API tokens, deleting files or establishing persistent footholds within otherwise secure network infrastructure.

IMPLICATION

Any account deploying AI systems for internal operations or other enterprise functions carries a new and largely unquantified risk. Standard cyber insurance applications are not asking the right questions, and most simply do not ask about it at all. CyRisk's AI underwriting tools, AI underwriting training, and vCAIO service are specifically designed to assess this.

03 WHY THE GUARDRAILS AREN'T ENOUGH

Project Glasswing's coalition and Anthropic's safety controls are the right response to what Mythos Preview revealed. They are also fighting a losing battle in the long run. The week after Anthropic's big announcements, OpenAI unveiled GPT-5.4-Cyber, specifically designed for cybersecurity tasks. Coincidence? Zeitgeist? Employees job hopping? Industrial espionage? It doesn't help that Anthropic's own security has performed the equivalent of a three stroke harakiri in less than a month, starting with the accidental leak of 3,000 unpublished assets, including the Mythos name. Then a second human error a few days later resulted in the accidental publishing of the entire 500K+ lines of source code for Claude Code. The final coup de grâce actually involved outsiders, but they employed the very low-tech TTP (that's Tactics, Techniques, and Procedures in cyber speak) of URL pattern guessing after gaining initial access from one of Anthropic's contractors. This enabled the unauthorized individuals to bypass Anthropic's protections and directly access the Mythos model. Anthropic acknowledged the incident but said the activity was limited to the third-party environment and did not extend to their internal production systems.

Given the excruciatingly poor performance of Anthropic's own internal security program, two questions come to mind: Are we still confident it will take many months for our adversaries to develop these capabilities? Are we confident the 50+ Glasswing coalition members will make enough progress shoring up our infrastructure and building defenses to stave off the Mythos grade cyber attacks to come?

IMPLICATION

Three structural limitations persist regardless of how well the coalition executes. Pricing these risks as if guardrails hold completely is working from an incomplete understanding of this new reality.

01 The Model-Weight & Training Problem

Anthropic's safety filters live in the deployment environment, not in the model weights. A stolen or replicated copy of a Mythos-class model carries no refusal behavior, no usage logging, and no coalition governance. Nation-state actors and well-resourced cybercriminal organizations have already demonstrated the capacity to exfiltrate frontier AI weights, steal probability distributions, and conduct feature distillation (e.g., DeepSeek's distillation of OpenAI's model). A growing class of adversarial techniques (prompt injection, targeted jailbreak methods, and the mechanisms above) is specifically engineered to bypass safety controls in deployed systems.

Enterprise AI deployments compound this further. A large language model running inside a policyholder's organization without prompt filtering or monitoring is not just a productivity tool; it is an uncontrolled entry point, like an insider threat that existing network security controls were not designed to catch.

02 The Volume Problem

Responsible disclosure assumes the CVE triage infrastructure can absorb what gets disclosed. That assumption breaks at Mythos-class discovery volume. The global Common Vulnerabilities and Exposures system, which was already struggling with new vulnerability discoveries before Mythos and is further hobbled by DOGE budget cuts, was not designed for a model that can produce dozens of critical findings in a single session.

When disclosed vulnerabilities outpace the capacity for validation and remediation, the gap between "Glasswing found it" and "your client patched it" *is* the attack window. A vulnerability that has been publicly disclosed but not yet addressed is not a secret. Patch velocity, already the primary loss predictor in the Glasswing Era, becomes the variable that determines whether a responsible disclosure becomes your client's claim.

03 The "Democratization" Problem

Glasswing constrains one model. It does not constrain the category. Widely available commercial and open source AI models (there is already an OpenMythos repo on GitHub), combined with accessible offensive security knowledge and the published Glasswing technical findings, have already lowered the attack skill floor to a degree that does not require Mythos-class capability. The Glasswing disclosure catalog, once public, hands both defenders and attackers an annotated map of what to look for and where. Even script kiddies now have a research agenda, and the AI tools to pursue it.

> *Even if Mythos were to remain locked down, the methodologies, training architectures, and broad capabilities will inevitably proliferate. The emergence of a Mythos-class AI in the hands of bad actors is, to use a very tired cliché, not a matter of if, but when.*

SECTOR SPOTLIGHT: FINANCIAL SERVICES

The four shifts above do not land evenly across all industries. Healthcare, manufacturing, utilities, transportation and IoT all have their thorny challenges and we will address each at a later date. But one sector sits at the intersection of every one of them, and as noted above, global regulators have recognized the threat and already started moving.

SECTOR SPOTLIGHT · MYTHOS MEETS FINANCIAL SERVICES

It's no wonder the April 7 emergency meeting at Treasury was not an isolated event, with bankers and finance ministers meeting around the globe, trying to get their heads around the Mythos question. A few years ago, Lloyd's published a Realistic Disaster Scenario that modeled the impact of a cyber attack on a major financial services payment system. Such an attack appears significantly less far-fetched today.

Banks have always been an attractive target. They are, after all, as Willie Sutton so eloquently pointed out, where the money is. But a significant portion of financial services also runs on a bedrock layer of legacy technology: decades-old core systems, mainframe code, COBOL, and aging integrations holding modern front-ends together. Much of that code predates modern secure coding practices and has never been subjected to the kind of adversarial scrutiny Mythos can now apply at machine speed and trivial cost. High-value targets sitting on exceptionally exposed software stacks puts financial institutions at the top of the Glasswing risk map, and global regulators know it.

Implication for cyber underwriters: *FI exposure now compounds across every shift in this brief: patch velocity on legacy systems, expert-attacker proliferation, supply-chain dependency depth, and AI deployment surface. Underwriters writing FI accounts should treat Glasswing-era controls as table stakes, not differentiators.*

04 WHAT THE INSURANCE MARKET CAN DO ABOUT IT

To address the emergence and proliferation of AI models like Claude Mythos capable of autonomous, machine-speed zero-day discovery and exploitation, the insurance market must shift from static, reactive models to dynamic, evidence-based underwriting, loss control and portfolio risk management. April 7, 2026 is the dividing line. The underwriters who move their risk signals now will be ahead of the first wave of post-Glasswing losses.

ACTION	RATIONALE	TIMEFRAME
Require patch-velocity & legacy-stack data in renewal submissions. Flag accounts with legacy OS or browser footprints for enhanced scrutiny.	Unpatched systems are now near-immediately exploitable post-disclosure; patch velocity is the #1 loss predictor in the Glasswing Era. Legacy OS and browser footprints carry compounding unpatched CVE density across every major platform Mythos probed.	IMMEDIATE
Incentivize phishing-resistant MFA. Mandate or discount for FIDO2 / WebAuthn standards.	Traditional MFA is increasingly vulnerable to AI-powered social engineering and real-time phish-kit automation; hardware-bound FIDO2 breaks the credential-theft loop that AI attackers rely on.	IMMEDIATE
Update applications for AI, open-source and supply-chain exposure. Add targeted questions and require SBOM analysis at submission and renewal.	Generic questionnaires miss the risks that now matter: enterprise AI deployments, open-source components (FFmpeg, OpenSSL, Linux kernel), and third-party dependency density. These are newly quantifiable, newly material risk factors.	ALL NEW & RENEWAL
Shift loss-control focus from "finding" to "fixing." Encourage automated remediation, micro-segmentation, and recoverability-centric controls.	Since AI makes vulnerability discovery cheap, the primary battleground is the unprotected window between disclosure and patching. AI-driven attacks also compress the lifecycle from compromise to disruption, so recovery capability now materially determines loss severity.	30 DAYS
Underwrite the AI and agentic systems themselves. Assess controls on policyholder LLM and agentic deployments.	As AI systems are deployed as privileged insiders across enterprises of all sizes, they become both prime targets and new attack surfaces. Insurers must come up to speed on AI risk and insist on appropriate mitigating controls, including prompt-firewall and monitoring requirements above retention thresholds.	60 DAYS
Update loss models and clarify policy language. Move toward affirmative AI coverage; revisit cat assumptions and reinsurance attachment points.	Existing models underestimate frequency for slow-patching accounts and simultaneous-attack correlation; "silent AI" exposure in current forms creates unintended accumulation. Explicit affirmative AI coverage reduces uncertainty for carriers, brokers, and insureds alike.	60 DAYS
Score the renewal book with the CyRisk Glasswing Risk Index. Baseline the portfolio against Glasswing-relevant signals.	Quantify exposure to the new vulnerability map *before* the first post-Glasswing losses arrive, so terms, pricing, and remediation requirements can move with evidence rather than in reaction.	60 DAYS

THE WINDOW TO ACT IS NOW.

Score your renewal book against the CyRisk Glasswing Risk Index before the first post-Glasswing losses arrive. CyRisk is ready to help carriers, MGAs, and reinsurers move from reactive to evidence-based risk decisions.

CONTACT
info@cyrisk.com

PHONE
(888) 478-1112

WEB
cyrisk.com

ABOUT CYRISK

CyRisk is a cyber risk analytics and data platform serving the cyber insurance market. Our platform delivers privacy, cyber, and AI technology risk intelligence (including tracking technology exposure, privacy policy gap analysis, AI risk indicators, and software stack fingerprinting), purpose-built for underwriting workflows. CyRisk provides both automated risk scanning and underwriter-ready reports for individual accounts and portfolio-level analytics. Our data is derived from continuous external monitoring, not self-reported questionnaires. For more information, or to discuss Glasswing risk scoring for your book, contact CyRisk at **info@cyrisk.com**.