HEXAWARE

AI & CYBERSECURITY – ANTHROPIC – PROJECT GLASSWING

Mythos, **Measured.**

What's real, what's overstated, and what to decide in next four months

AUTHOR

Satyajith Mundakkal

CTO, Hexaware

Published

May 2026

12 min executive skim – 60 min full report

THE SHORT VERSION

- **What it is.** Claude Mythos Preview is Anthropic's restricted frontier model, available through Project Glasswing to 12 launch partners plus over 40 additional organizations. Not generally available and not buyable.
- **What's validated.** AISI confirmed Mythos as the first frontier model to complete a 32-step simulated network attack end-to-end. Mozilla shipped 271 Firefox 150 fixes traceable to one Mythos evaluation. The body of this brief unpacks the methodology, the limits, and the items the headlines overstated.
- **What a recipient actually got.** On May 11, curl maintainer Daniel Stenberg published the first inside-the-program account of a Mythos scan. Of five findings the model labeled "Confirmed security vulnerabilities," exactly one survived curl's security review as a real CVE. The other four were false positives on documented behavior or non-security bugs. Two side notes worth keeping: Stenberg never got direct access — a third party ran the scan and forwarded the report — and the lower-confidence findings in the same report held up well. The hype was overstated; the category isn't.
- **What changed this week.** On May 11/12, Google's Threat Intelligence Group disclosed the first case it has identified of a zero-day exploit it believes was developed with AI — a 2FA bypass attributed to a cybercrime group, not Mythos and not Gemini. AI-assisted vulnerability discovery and exploit generation is now in live adversary workflows, even if Mythos-class capability is not yet broadly available.
- **What to do.** Measure and drive down mean time from validated finding to verified production fix. The discovery side is commoditizing; remediation is now the bottleneck and the differentiator.

Bottom line up front

A real jump, oversold.

Claude Mythos Preview is a real capability jump in AI-assisted cybersecurity. The marketing around it overstates the magnitude in places, and several flagship findings have softer foundations than the headlines suggested. What's harder to dismiss is the trajectory the announcement reveals.

Two external validations make the capability credible. AISI — an independent UK government evaluator — confirmed Mythos as the first frontier model to complete a 32-step simulated corporate network attack end-to-end. Mozilla, working inside a real release process, shipped 271 vulnerability fixes in Firefox 150 traceable to a single Mythos evaluation. Both validators were careful to note that scaffolding (harnesses, test infrastructure, expert workflow) does much of the work. The model is necessary but not sufficient. The detailed methodology behind each headline figure, and the limits the evaluators named, are unpacked in Section 02.

The strategic implication has been underplayed. AI is collapsing the time and skill required to find vulnerabilities, while remediation capacity has not moved. Over 99% of Mythos's candidate vulnerability findings remained unpatched as of Anthropic's April red-team write-up. The interval from vulnerability disclosure to confirmed exploitation has compressed from roughly 2.3 years in 2018 to roughly 20 hours by April 2026. The question for the next four months isn't whether to buy Mythos — you can't. It's whether your remediation pipeline is built for machine-pace discovery on the other end, because AI-assisted vulnerability discovery and exploit generation has entered live adversary workflows, and comparable Mythos-class capability may become accessible on a months-not-years timeline. This brief treats 6–18 months as a scenario window for planning, not as a verified Anthropic forecast.

Section 01 · Disambiguation

Three things, often confused.

Public conversation about Mythos has merged three distinct things. Clarity here saves time later — in board discussions, in vendor negotiations, and in reading the press.

When someone says "Mythos can do X," the first useful question is whether they mean the underlying model, the consortium organized around it, or the shipping product available to enterprise customers today. The answer to each is different, and the strategic implications diverge.

TERM	WHAT IT IS	STATUS	WHAT IT ISN'T
Claude Mythos Preview	Anthropic's frontier model, gated for research and partner use. General-purpose, with unusually strong cybersecurity, coding, reverse-engineering, and agentic capabilities. Discovered, per Anthropic, during testing — not designed-for.	NOT FOR SALE No general commercial availability has been announced. Available only through Project Glasswing partners.	Not a scanner. Not a security product. Not a "buy now" item, regardless of vendor pitches that imply otherwise.
Project Glasswing	The restricted-access program around Mythos. 12 launch partners (AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, Linux Foundation, Microsoft, NVIDIA, Palo Alto Networks, Anthropic), plus over 40 additional organizations. $100M usage credits committed; $4M to open-source security.	RESEARCH PREVIEW Active since April 7, 2026.	Not a standards body. Not a regulatory framework. Public materials do not disclose a formal enforcement mechanism or public accountability process for partner obligations.
Claude Security *(previously Claude Code Security)*	A productized Claude Enterprise workflow that reasons over code, validates findings, and proposes patches for human approval, with patch handoff into Claude Code on the Web. AI-native SAST-style tooling, but with deeper code reasoning than pattern-match scanners.	AVAILABLE Public beta for Claude Enterprise customers.	Not the Mythos model. Narrower in scope. Does not perform runtime detection, intrusion response, or live attack simulation.

The conflation has consequences. Vendors pitching "Mythos-class" capabilities are usually referring to Claude Security (previously known as Claude Code Security), which is real but narrower. Analysts panicking about an AI cyber weapon are usually referring to the Mythos model itself, which most enterprises will not see unless they are admitted to Glasswing or Anthropic announces a broader release. Boards asking whether "we have access to Mythos" typically mean Project Glasswing — and unless the organization is one of those partners, the answer is no.

Anthropic is not alone in occupying this layer. OpenAI ships Codex Security — the productized version of its Aardvark agent — to ChatGPT Enterprise and Business customers, and runs a parallel gated cyber access program (Trusted Access for Cyber) around GPT-5.5 and the more permissive GPT-5.5-Cyber for verified defenders. In May 2026 OpenAI also packaged these pieces together under *Daybreak*, a Codex-Security-plus-GPT-5.5 cyber-defense initiative with its own partner list (Cloudflare, Cisco, CrowdStrike, Palo Alto Networks, Oracle, Akamai, and others). It is the most direct structural answer to Project Glasswing in the market. Google ships Gemini Code Assist Enterprise with a security analysis extension, alongside its Big Sleep vulnerability-discovery agent and CodeMender patching tool. The three-layer confusion above — model, gated program, productized capability — repeats inside each of those ecosystems. The vendor names change. The procurement question does not.

Timeline of public events

26 MAR 2026

The accidental leak

Anthropic CMS misconfiguration exposes ~3,000 unpublished assets, including draft material about an unreleased model named Mythos. Cybersecurity stocks fall within 24 hours.

7 APR 2026

Official announcement

Anthropic announces Claude Mythos Preview and Project Glasswing. ~250-page system card published alongside. White House briefed.

13 APR 2026

Schneier responds

Bruce Schneier publishes the first sustained skeptical commentary, calling it "very much a PR play by Anthropic — and it worked."

MID-APR 2026

Independent evaluations

UK AI Security Institute (AISI) and AISLE, an AI cybersecurity startup, publish independent assessments. AISI confirms a "step up"; AISLE argues the moat is the system, not the model.

21 APR 2026

Firefox 150 ships

Mozilla releases Firefox 150 with fixes for 271 vulnerabilities found during the initial Mythos evaluation. First real-world operational result at scale.

MAY 2026

Mozilla explains the harness

Mozilla publishes the engineering write-up describing fuzzing infrastructure, triage workflow, rollup CVEs, and release process. The "scaffolding, not magic" account.

11 MAY 2026

The first inside-the-program review

curl maintainer Daniel Stenberg publishes what a Mythos scan actually produced when sent to one of the program's open-source recipients. Of five findings the model called "Confirmed security vulnerabilities," curl's team kept one — a low-severity CVE shipping with curl 8.21.0 in June. The scan was run by a proxy after direct access never showed up. (Source)

Why this is more than vendor marketing

Two facts elevate Mythos above ordinary product-launch coverage. First, two external technical validations — AISI (a UK government body, fully independent of Anthropic) and Mozilla (one of the world's most security-rigorous open-source organizations, operating inside a real release process) — confirmed material capability gains under their own methodology and harnesses. Neither is in the business of validating Anthropic's marketing.

Second, Anthropic chose *not* to make the model generally available. That decision costs the company revenue and would be irrational as pure marketing strategy. Restriction can be reframed as scarcity-driven hype, and skeptics including Schneier have argued exactly that. But the more defensible reading is that Anthropic has assessed the offensive risk as real enough to absorb the commercial cost of withholding.

> *Skepticism about specific Mythos claims is warranted. Skepticism that something significant happened is harder to defend.*

What follows examines the capability claims with that posture: take the evidence seriously, separate independently validated findings from vendor extrapolation, and focus on the strategic shift the technology forces — which is larger than the specific model and will outlast its release cycle.

Section 02 · The evidence, sorted

What it actually does — by source quality.

Capability claims about Mythos have spread faster than the evidence behind them. Some figures come from independent technical evaluators with no commercial interest in the outcome. Some come from Anthropic's own materials. And some of the most-cited bugs turn out, on closer inspection, to be less devastating than the headlines suggested.

Three tiers separate cleanly, each on its own terms. Tier 1 is what holds up under independent scrutiny. Tier 2 is what's plausible but rests on vendor methodology. Tier 3 is what Mythos demonstrably could not do, at least at the point of evaluation.

TIER 1 · INDEPENDENTLY VALIDATED

What two outside evaluators confirmed.

UK AI Security Institute · Mozilla Corporation

32

steps autonomously completed
First frontier model to solve AISI's "The Last Ones" corporate network simulation end-to-end. Succeeded in 3 of 10 attempts; averaged 22 of 32 steps versus 16 for the prior model.

271

Firefox vulnerabilities patched
Mozilla shipped 271 security fixes in Firefox 150 traceable to a single Mythos evaluation. April releases totaled 423 security bug fixes, with the Mythos-aided pipeline contributing the largest share.

73%

expert-level CTF success
Capture-the-flag performance under AISI evaluation. Continued improvement on multi-step cyber-attack simulations compared to the prior frontier generation.

These are the load-bearing claims. AISI is a UK government body that has been tracking AI cyber capabilities since 2023 and builds progressively harder evaluations specifically to test frontier models. Mozilla has one of the most security-rigorous engineering cultures in commercial open source, and shipped the fixes through a normal release cycle. Mozilla's own write-up notes that many findings were grouped into rollup CVEs and severity classifications rather than published as one standalone CVE per bug, so the 271 figure refers to fixed vulnerabilities, not 271 standalone exploitable zero-days. Neither evaluator has a commercial interest in inflating Anthropic's claims.

The pattern in both evaluations is the same: Mythos represents a step up from the previous frontier, not a discontinuity. AISI describes the gain as significant but continuous with prior improvement trajectories. Mozilla characterized the model as comparable to elite human security researchers — strong language, but framed against a baseline of human capability, not against magic.

Sources · UK AI Security Institute, *Our evaluation of Claude Mythos Preview's cyber capabilities* (April 2026) · Mozilla, *The zero-days are numbered* (May 2026) · Mozilla Hacks, *Behind the Scenes Hardening Firefox with Claude Mythos Preview* (May 2026) · Firefox 150 release notes (21 April 2026).

TIER 1 · INSIDE THE ACCESS PROGRAM

Inside the access program: the curl scan.

Daniel Stenberg · curl maintainer · 11 May 2026

5 → 1

"confirmed" findings vs. confirmed by maintainer
Mythos labeled five findings as "Confirmed security vulnerabilities." curl's security team verified one. Three of the other four were false positives on documented API behavior; the fifth was a non-security bug.

178K

lines of C scanned
Scan covered curl's `src/` and `lib/` directories at a specific git commit. Methodology disclosed in the report: subagent-driven parallel file reads with source-level re-verification in the main session. No traditional SAST involved.

1

CVE pending in curl 8.21.0
The single confirmed vulnerability ships as a low-severity CVE with curl 8.21.0 in late June. Details embargoed until release. The roughly twenty other findings — flagged as bugs, not vulnerabilities — held up well under review.

On May 11, Daniel Stenberg — who has maintained curl since 1998 — published an account of what a Mythos scan actually delivered to a project outside the 12 named launch partners. It's the first time anyone on the receiving end of the program has shown their work in public.

What he got is more interesting than the simple version of the story. Access for open-source projects was routed through the Linux Foundation, with Alpha Omega administering distribution. Stenberg signed the contract. Weeks went by, and the access never arrived. Eventually someone else — someone who did have access — offered to run the scan for him and send the report. He accepted. So the "first inside-the-program review" was actually a third party running Mythos against curl and handing curl's maintainer the result. That handoff isn't in any of the published Glasswing materials.

The report disclosed its own method up front: a human driver using Mythos in subagent mode for parallel file reads, with every candidate finding re-checked against the source in the main session. No traditional SAST tooling involved. The scan covered 178,000 lines of C in curl's `src` and `lib` directories at a specific commit.

Mythos labeled five findings as "Confirmed security vulnerabilities." That word *confirmed* is doing a lot of work — it's the model's confidence, not anyone else's. Stenberg's security team went through the five and kept one: a low-severity CVE that ships with curl 8.21.0 in late June. Three of the other four were false positives on behavior that's documented in curl's own API. The fifth was a real bug, but not a security one.

The roughly twenty other findings in the report — flagged as bugs rather than vulnerabilities — held up much better. "Barely any false positives," Stenberg wrote. So Mythos didn't have a noise problem in general. It had a noise problem specifically in the tier the press would have quoted: the "confirmed" headline number.

Stenberg's own read separates the marketing from the technology. Over the previous eight to ten months, AISLE, Zeropath, and OpenAI's Codex Security had already scanned curl and contributed something on the order of two to three hundred merged fixes and a dozen-plus CVEs between them. By the time Mythos got there, the easy material was gone. Even allowing

for that, Stenberg writes that he sees no evidence Mythos finds anything to a "particularly higher or more advanced degree" than tools that came before it. The hype, in his words, was "primarily marketing." And yet — same post, same author — "AI powered code analyzers are significantly better at finding security flaws and mistakes in source code than any traditional code analyzers did in the past." Both things are true. The category is real. The leap inside the category is smaller than the press release suggested.

One scan, one codebase, one maintainer doing the verification. Don't extrapolate this into a Mythos false-positive rate. What it does do is answer, in part, the question Schneier kept asking and Anthropic kept declining to publish: when you actually look at what the model called "confirmed," what fraction holds up? Five-to-one is one data point, not a denominator. A second project publishing its numbers would change the picture either way. Until then, treat the 5 → 1 as a reason to keep the verification step expensive — and a reason to be careful about anyone closing tickets on a scanner's confidence alone.

Source · Daniel Stenberg, *Mythos finds a curl vulnerability*, daniel.haxx.se (11 May 2026).

TIER 2 · VENDOR CLAIMS, WITH CAVEATS

Where the headlines come from.

Anthropic system card · Project Glasswing announcement

The "thousands of zero-day vulnerabilities" figure that drove much of the press coverage is real in the sense that Anthropic published it. It is also a sampled-validation projection. The underlying methodology: human expert contractors manually reviewed 198 of Mythos's flagged findings. Anthropic reports exact severity agreement with the model in 89% of those cases, and agreement within one severity level in 98% of cases. That agreement rate was then scaled to the full corpus of model-generated findings to produce the headline figure.

This is a defensible approach. It's also worth understanding for what it is: a sampled-validation projection, not a count of independently confirmed bugs. The denominator — how many findings the model produced in total, including false positives that didn't survive triage — has not been disclosed publicly. Schneier's central skeptical point lands here: without a clear false-positive rate, the headline figure overstates the practical signal-to-noise ratio of running Mythos against unfamiliar code.

Other vendor claims warrant similar treatment. Anthropic reports that Mythos found exploitable conditions in every major operating system and web browser. The Red Team report describes successful reverse engineering of stripped binaries to reconstruct source-equivalent representations. AISI separately confirmed multi-stage autonomous network attacks in controlled environments. Each of these is consistent with a genuine capability jump. None of them rises to the level of "an AI system that can break arbitrary modern systems on demand."

Sources · Anthropic, *Project Glasswing* announcement (April 2026) · Anthropic Frontier Red Team, *Claude Mythos Preview* (April 2026) · Schneier on Security, *On Anthropic's Mythos Preview and Project Glasswing* (April 13, 2026).

TIER 3 · WHAT IT COULD NOT DO

The ceiling, where it has been tested.

Anthropic re-analysis · AISI environmental caveats

Three points push back against the more excitable readings of the announcement.

First, Mythos could not weaponize several of its most-cited findings against hardened targets in the form initially presented. Anthropic's own deeper analysis of the celebrated FFmpeg bug — a 16-year-old flaw that survived roughly five million automated fuzzing tests — concluded that the vulnerability was not of critical severity and would be challenging to exploit in practice. On Linux, the picture is mixed: Anthropic reports that many remotely triggerable out-of-bounds kernel findings could not be turned into working remote exploits, particularly against modern mitigations — address-space layout randomization, control-flow integrity, kernel page-table isolation. But Anthropic also reports successful local privilege-escalation exploit chains, including nearly a dozen examples where two-to-four kernel vulnerabilities were chained to reach root access. The right conclusion is not "Linux findings were non-exploitable"; it is that exploitability varied by target, mitigation stack, and chain completeness.

Second, AISI was explicit that its test environments did not include active defenders. The 32-step network simulation was a vulnerable range — designed to be exploitable so the evaluation could measure progress. AISI's own framing was that Mythos would be effective against systems with weak security posture, but the evaluation does not establish what happens when the model is pointed at a hardened production environment with EDR, segmentation, anomaly detection, and a competent incident response team. That test has not been publicly run.

Third, AISLE, an AI cybersecurity startup, conducted a separate experiment in which small, cheap, open-weight models were pointed at the same vulnerability locations Mythos had publicly showcased. Several of them — including a 3.6 billion parameter model running at roughly eleven cents per million tokens — recovered substantial portions of Mythos's analysis when the code path was isolated and scoped for them. AISLE was careful about the caveat: the smaller models needed hints and contextual prompts that Mythos was not given. So the comparison is not apples-to-apples. But the finding has a real implication, which AISLE stated directly: there is no stable "best model for cybersecurity," and the capability frontier is jagged. Much of the operational advantage comes from the system built around the model — the harness, the test infrastructure, the workflow — rather than from raw model intelligence.

Sources · Anthropic system card re-analysis (April 2026) · AISI evaluation environmental notes · AISLE, *AI Cybersecurity After Mythos: The Jagged Frontier* (April 2026).

THE SCAFFOLDING POINT

Mozilla's engineering write-up is the most useful corrective to the "magic AI hacker" framing. The 271-bug Firefox result did not come from running the model against the Firefox source and waiting for output. Mozilla built a project-specific agentic harness on top of existing fuzzing infrastructure. The harness generated reproducible test cases. It dynamically tested whether the model's hypotheses about bugs held up. Jobs were parallelized across ephemeral virtual machines. Findings were deduplicated, triaged through a normal bug-tracking process, patched, regression-tested, and integrated into a normal release.

Mozilla's own characterization: a stronger model was necessary but not sufficient. The capability gain came from "steering, scaling, stacking, reproducing test cases, filtering noise, triage, and release integration." That sentence — not the 271 number — is the most important operational claim in the Mozilla writeup.

This matters strategically because it locates the moat. If the value of Mythos-class capability lives mainly in the model weights, then control of access to weights is the central governance lever. If much of the value lives in the systems built around the model, then the governance question is broader and the competitive question is different. Both readings have implications later in this brief.

Why the evidence tiers matter

The reason for sorting capability claims by source quality is not pedantic. It changes the strategic calculation. If the headline "thousands of zero-days across every major OS" were a confirmed independent finding, the appropriate response would be near-panic. The actual evidence — strong capability gains validated independently, scaled-up headline numbers that depend on vendor methodology, and meaningful operational limits at the hardened-target frontier — supports a different posture: serious attention, deliberate preparation, no rushed decisions.

The other reason is that vendor pitches in the next twelve months will reference Mythos as justification for products that may not warrant the inference. A SAST vendor claiming "Mythos-class detection" is making a Tier 2 claim at best. A pentesting firm claiming "AI-augmented red team services" is operating in Tier 1 territory only if they can demonstrate evaluation results with the rigor of AISI or Mozilla. The tier framework gives a CXO a quick test for separating the credible from the aspirational.

IF YOU DON'T HAVE A CISO IN THE ROOM

The practical test for any vendor claim: ask which tier the evidence lives in. Independent technical evaluation under named methodology is Tier 1. Vendor-reported metrics without disclosed denominators are Tier 2. Anything that hasn't been run against a hardened target with active defense is Tier 3 — interesting but not yet proof of operational capability.

Section 03 · The core argument

The finding-to-fix gap.

Mythos has decoupled vulnerability discovery from vulnerability remediation. That decoupling, not the model itself, is the strategic problem.

For most of the history of computer security, vulnerability discovery and vulnerability remediation lived in roughly the same time domain. Researchers found bugs at a human pace. Maintainers patched bugs at a human pace. The disclosure window — typically 90 days from notification to public release — was calibrated to the throughput of both sides. That equilibrium worked because the constraint on both finding and fixing was the same: skilled human attention.

AI changes one side of that equation and not the other. Mythos and its successors can find vulnerabilities at machine pace. The pipelines that absorb, validate, prioritize, patch, regression-test, deploy, and verify those fixes remain firmly on human pace. The result is a widening gap — and that gap, not the model itself, is the strategic problem.

Disclosure-to-exploitation timing is compressing

Directional three-point timeline. Not a measured series, not an operational SLA, not a Project Glasswing evaluation result.

Sources · IAPS / zerodayclock data, drawing on CISA KEV, VulnCheck KEV, and XDB feeds, as summarized in IAPS, *Mythos and the Evolving Cyber Landscape* (April 2026). The 2018 figure (~2.3 years) and 2025 figure (~23 days) reflect that aggregated data. The April 2026 figure (~20 hours) reflects reported disclosure-to-exploitation timing during the Mythos period; it is not a direct Mythos, Glasswing, or AI-attribution measurement, and is not a Project Glasswing evaluation result. Treat the chart as directional rather than as a precise series.

That compression is the operational reality. An organization that took six weeks to patch in 2018 had nearly two years of slack. An organization that takes six weeks to patch in 2026 has none. The defensive timing assumption built into most enterprise vulnerability management programs —

quarterly scans, monthly patch windows, ticketed backlogs, change advisory boards meeting every two weeks — was calibrated to a threat landscape that no longer exists at the high end.

The diffusion clock

Mythos's restriction inside Project Glasswing buys defenders time. The question is how much, and the honest answer is uncertain. Public commentary supports a months-not-years planning posture, while policy analysis around frontier-model lag suggests adversaries may not remain far behind indefinitely. This brief therefore treats 6–18 months as a scenario window for executive planning, not as a measured forecast and not as an Anthropic-published estimate.

PLANNING ASSUMPTION	ALEX STAMOS · RSA 2026	FRONTIER LAG
6 – 18 months	**months, not years**	**illustrative pattern**
Scenario window used in this brief for executive planning. Informed by public expert commentary and frontier-lag analysis; not an Anthropic-published forecast.	"Every nineteen-year-old in St. Petersburg." Stamos's framing for how quickly capability moves from frontier lab to broadly accessible tooling outside Western institutional control.	Approximate gap between US frontier models and the leading capability available outside Western alliance jurisdiction, narrowing over time. Open-weight lag has compressed substantially over the past two generations, though exact figures depend on benchmark and task.

The clock is a risk-planning device, not a benchmark.

Three practical implications follow from holding the six-to-eighteen-month window loosely rather than tightly. First, the defender's head start is real but bounded. Second, whatever an organization does not patch during that window may be exploitable afterward at a meaningfully lower attacker skill threshold than before. Third, the patching that matters happens at the upstream layer — the operating systems, browsers, kernels, and shared libraries that compose the application stack — because that is where the capability advantage gets converted into widespread real-world safety, or fails to.

Which code goes first

There is a corollary to the diffusion clock rarely stated explicitly. When AI vulnerability discovery operates at machine pace, the order in which different codebases get scanned is not random. Source-code availability is not the only ordering variable — deployment scale, exploitability, internet exposure, patch latency, bounty value, and strategic value all matter to attackers and defenders alike. But source availability is the variable that most directly lowers the cost of model-driven inspection, and the exposure tiers fall out cleanly along it.

The mechanism is simple and asymmetric. Source-code-available projects can be scanned by any sufficiently capable model without special access. Mozilla's 271-fix Firefox 150 result is the operational proof of what that means at scale. Open-core projects expose their foundation fully; the enterprise commercial layer adds some marginal protection but rests on a base that does not. Closed-source software is the least directly exposed to AI source-code analysis, though it is not immune — Anthropic's own report explicitly lists reverse-engineering of stripped binaries among Mythos's capabilities, which raises the cost of attack on closed systems substantially without eliminating it.

The closed-source framing also understates exposure in a second direction, and this one matters more. Most modern products in the Less exposed tier are built atop open-source dependencies — Linux distributions, JVMs, OpenSSL, glibc, log4j, Spring, sqlite, libcurl — that already sit on the Most exposed tier and inherit its full scanning surface. SAP and Oracle do not write their own TLS stacks. Banking core systems run on shared kernels. Even classified defense applications link against the same C runtime libraries that ship in every Linux distribution. A vulnerability discovered in a foundational library propagates into every closed-source product that ships with it, which is essentially all of them. The proprietary application layer is shielded; the substrate underneath is not, and the substrate is where the most severe and most reusable bugs tend to live.

One subset of the Less exposed tier is a partial exception to the dependency-inheritance point, not because these systems are immune but because their substrates are often less exposed to commodity open-source source-code scanning. Mainframe and midrange systems — z/OS, IBM i,

COBOL and RPG codebases, AIX, HP-UX, Solaris, and the transaction-processing platforms underneath airline reservation systems — are categories where the substrate itself is typically proprietary. An airline running Sabre on IBM's TPF operating system has limited Linux dependency to inherit exposure from. A bank running core ledgers on z/OS with DB2 and CICS is not, for those workloads, linking against glibc.

The combination of closed application code, closed operating system, niche programming languages, and limited model training data on those languages produces a genuinely harder target for AI source-code analysis than the rest of the tier. That is not the same as "safe." Modernization adds OSS surface — z/Linux partitions, Java on z/OS, Python on IBM i, OpenSSL ports — and decades of unscanned legacy code in the financial and aviation systems is itself a sitting risk profile that has been deferred rather than addressed.

For the moment, the mainframe layer is plausibly where AI security tooling is least productive — a reading consistent with the dependency-inheritance pattern described above, but author analysis rather than a measured finding — which happens to also be the layer running the most consequential systems in banking, defense, and airline operations.

The pattern explains part of why Project Glasswing's partner composition makes the strategic sense it does. The closed-source vendors in the consortium are buying preemptive defense before the cost of binary analysis falls further. The Linux Foundation is in because shared open-source code has no source-code gate to defend in the first place. Over the next twelve months, open-source maintainers will absorb the largest share of AI discovery output, and the funding asymmetry between what runs the internet and who pays to maintain it becomes the central operational risk at the upstream layer.

The ninety-nine percent problem

The number worth dwelling on isn't the headline capability figure. It's what's happened to the findings since.

99%+

of the candidate vulnerabilities Mythos has identified through Project Glasswing remain unpatched as of Anthropic's April red-team write-up.

"Candidate" here matters: the public record supports a large disclosure pipeline, not an independently verified count of thousands of deployable exploits. Anthropic and its partners now possess a large corpus of candidate high- and critical-severity vulnerabilities, many of which appear serious enough to require professional triage and coordinated disclosure. The public evidence supports a real surge in discovery output; it does not independently confirm every candidate as a working exploit. The mechanism for converting that knowledge into deployed fixes has barely begun to move. The bottleneck isn't detection, or even prioritization. It's the entire downstream pipeline: coordinated disclosure to maintainers, patch authoring, regression testing, release management, deployment to running fleets, and verification that the fix actually closed the underlying weakness.

Anthropic has hired contractors specifically to triage the disclosure pipeline at Mythos's output rate. That hiring decision is itself an admission. The responsible disclosure infrastructure that has served the industry since the early 2000s was not built for machine-scale finding. The ninety-day default disclosure window was calibrated to the throughput of a human reviewer working alongside a human maintainer. It does not survive contact with a model that can generate plausible vulnerability candidates faster than the contracted human triagers can sort them.

Forrester's analysts have stated the implication directly: Project Glasswing breaks the existing vulnerability management playbook. Periodic scanning, monthly patch cycles, the CVE-to-Patch-Tuesday rhythm, and static asset inventories were all components of a system designed for the prior equilibrium. Each of them is now strained, and at the high end of attacker capability, several are functionally obsolete.

What this means for value

The strategic shift is best understood as a migration of value within the security function.

Before AI-scale discovery, the scarce thing was knowing what was wrong. Scanners, pentests, bug bounties, and red teams were all valuable because they generated signal about unknown weaknesses. The bottleneck was finding, and the budget followed accordingly.

After AI-scale discovery, the scarce thing is everything downstream of finding. Validation that a candidate finding is actually exploitable. Prioritization across thousands of plausible reports. Authoring of patches that don't introduce regressions. Testing that confirms the patch holds under realistic conditions. Deployment infrastructure that can push fixes through change management at a cadence the threat landscape now requires. Audit trails that demonstrate to regulators and insurers that the loop was actually closed.

> *The bottleneck has moved. The budget has not yet followed.*

Most enterprise security organizations are still resourced for the prior equilibrium. They have invested in scanning, threat intelligence feeds, and EDR platforms that detect known patterns. They

have invested far less in the unglamorous downstream machinery — change management automation, regression test suites for security patches, rapid deployment pipelines for OS-level updates, compensating controls that can hold the line when patching is delayed. That mismatch is the central planning problem for the next six to eighteen months.

> THE SINGLE METRIC THAT MATTERS NOW
>
> *Mean time from validated finding to deployed fix, measured across the production fleet — not from finding to ticket-closed, not from finding to patch-released, but from finding to verified-in-production. Most organizations cannot currently calculate this number with confidence. The first task is to be able to measure it. The second is to drive it down by an order of magnitude.*

Specific Mythos claims will move around in the coming quarters. Headline figures will be re-examined, several flagship findings will be partially walked back, and new evaluations will surface that test the model against hardened production environments rather than vulnerable ranges. That churn is the normal arc of any capability announcement. What it doesn't change is the decoupling underneath. The organizations that come through the next eighteen months in good shape will be the ones who treated remediation throughput as the problem, not vendor selection.

Section 04 · Governance reality

Who has access, and what it tells you.

Project Glasswing presents itself as a defensive cybersecurity research preview. The framing is accurate but incomplete — and the incompleteness matters for any organization trying to read the program's actual strategic purpose.

HOW TO READ THIS SECTION

The claims that follow live in three different evidence categories, and that distinction changes how to weigh them. **Company-confirmed:** the launch partners, the 40-plus-organization extended group, the $100M usage credits, the $4M open-source donations, the defensive-use framing, and the per-token pricing — all stated by Anthropic directly. **Press-reported:** Bloomberg and Euronews on bank testing and the Treasury-convened April meeting; press accounts of White House and Treasury-level engagement; Bloomberg's report of unauthorized third-party access via a contractor. **Author inference:** the read of Glasswing as US national-security industrial policy, and the description of administration influence over partner-list expansion as an effective veto. Each paragraph below should be read against the category it sits in.

The surface description is straightforward. Twelve launch partners plus over 40 additional organizations have access. Up to $100 million in model usage credits has been committed. An additional $4 million has gone in direct donations to open-source security organizations including Alpha-Omega, OpenSSF, and the Apache Software Foundation. Partners are subject to Anthropic's Usage Policy and are expected to use Mythos for defensive purposes only. The model is priced at $25 per million input tokens and $125 per million output tokens — roughly five times the rate of Anthropic's standard frontier model.

The more interesting story is in the composition of the partner list, the inclusion patterns it reveals, and the gap between stated governance and actual enforcement.

Beyond the named partners

The figure above describes the formal Glasswing partner footprint. The actual access footprint is broader. The items in this section come from press reporting rather than from Anthropic, and the company has not publicly confirmed the additions that go beyond the launch-partner list; readers should treat them as directionally credible rather than company-verified.

UK AISI clearly had evaluator access close to or before the public announcement — it published its Mythos evaluation on April 13, six days after launch, and the depth of the evaluation strongly

suggests preceding access, though Anthropic has not publicly described the exact access pathway. Press reporting separately indicates the US Commerce Department's CAISI was testing Mythos, but the timing and access mechanism should be treated as press-reported unless Anthropic or CAISI confirms it directly. Mozilla had "early access" beginning in February, first with Opus 4.6, then with Mythos for the Firefox 150 work that produced 271 fixes; Anthropic has not clarified whether this counts as part of the extended partner group or as a separate arrangement that predates the program.

Bloomberg and Euronews have reported that Goldman Sachs, Citigroup, Bank of America, and Morgan Stanley are testing Mythos following a closed-door April meeting convened by Treasury Secretary Scott Bessent. That extends the banking-sector access well beyond the single named partner, JPMorgan Chase.

A fifth pathway runs through the Linux Foundation, with Alpha Omega administering distribution to open-source projects. On paper, it looked like the same arrangement as the other partners — sign a contract, get access. The 11 May curl disclosure shows it didn't always work that way. Daniel Stenberg signed the contract and never got hands-on access. Someone else with access ran the scan for him and forwarded the report. That handoff is real, and it isn't documented anywhere in the published Glasswing materials. If you're sizing up what Mythos can do for your stack in a procurement conversation, the distinction between direct access and proxied access is worth surfacing — they're not the same thing.

And on the same day Glasswing was announced, a private Discord group focused on tracking unreleased AI models reportedly gained ongoing access through a third-party Anthropic contractor, using a mix of credentials from a contractor employee and URL pattern enumeration informed by an earlier Mercor breach. Bloomberg corroborated the report with screenshots and a live demonstration; Anthropic confirmed it was investigating "a report claiming unauthorized access to Claude Mythos Preview through one of our third-party vendor environments" and said there was no evidence its own systems were affected. Treat this as press-reported, not company-confirmed, and as describing the third-party vendor environment, not Anthropic's core infrastructure. A separate claim by a ShinyHunters impersonator the following day — circulating AI-fabricated dashboard screenshots — was quickly debunked by security researchers and is unrelated to the Bloomberg-sourced report.

Either way, the publicly named partner list understates the actual access footprint, and the unauthorized-access channel has not been publicly confirmed as closed.

Sources for this subsection · UK AISI Mythos evaluation (April 13, 2026) for pre-deployment evaluator access · Mozilla, *Behind the Scenes Hardening Firefox* for the Opus 4.6 / Mythos sequencing on Firefox 150 · Bloomberg, *Bessent, Powell Summon Bank CEOs to Urgent Meeting Over Anthropic's New AI Model* (April 10, 2026; paywalled) for the Treasury-convened bank-CEO meeting · CNBC follow-up on the Powell-Bessent bank-CEO meeting (April 10, 2026) · Euronews on the Bessent meeting and European banking-sector context (April 22, 2026) · TechCrunch on the unauthorized-access report. Specific quoted figures attributable to each publisher; treat publisher-attributed claims as press-reported rather than company-confirmed.

The pattern in who got selected

Anthropic's stated criterion for Glasswing inclusion is "organizations that build or maintain critical software infrastructure." That criterion is broad — it could plausibly apply to thousands of organizations worldwide, including major European industrial software firms, Asian semiconductor and telecommunications companies, and Indian IT services giants that maintain code running in production at scale across multiple continents.

The criterion actually applied was narrower, at least among the publicly named partners. Every named launch partner is headquartered in the United States. The composition maps cleanly onto the technology stack underneath the US federal government, US defense supply chain, and US-critical financial infrastructure. The forty-plus additional organizations have not been publicly disclosed; conclusions about the full access footprint's geography should therefore be read as based on the named launch list, not on the complete partner set.

HYPERSCALERS *FedRAMP High and DoD authorized cloud workloads*	Amazon Web Services · Microsoft · Google
ENDPOINT & NETWORK SECURITY *Major federal contracts; deployed across DoD and intelligence*	CrowdStrike · Palo Alto Networks
NETWORK & SILICON *Embedded in federal and defense supply chains*	Cisco · Broadcom · NVIDIA
CRITICAL FINANCIAL CLEARING *Primary Treasury clearing relationships*	JPMorgan Chase
ENDPOINT DEVICES *Secure devices in federal and executive use*	Apple
OPEN-SOURCE FOUNDATION *Maintains code that runs across federal and defense infrastructure*	Linux Foundation

The absences are as informative as the inclusions. No major European industrial software vendor. No Asian semiconductor or device manufacturer. No Indian IT services firm — despite the fact that companies of that profile maintain enormous amounts of production code for global clients. No federal systems integrator or services consultancy of the Accenture or Deloitte type, though this last absence has a simpler explanation: services firms typically do not own product source code in the way Mythos's discovery workflow is optimized for. The pattern fits a specific selection criterion: product vendors whose code composes the US federal, defense, and critical financial technology stack.

Why the Linux Foundation fits the pattern

The Linux Foundation's inclusion looks at first like a complication to that read. Linux is global open source. Patches to the kernel benefit every operating system distribution worldwide, including those deployed in jurisdictions that the rest of the partner list seems carefully designed to exclude. If Glasswing were narrowly about protecting a US technology stack, Linux would seem to leak the benefit.

The logic resolves once the diffusion clock from the previous section is brought into the analysis. Linux kernel source is already public. Any sufficiently capable adversary AI model can be pointed at the same code and asked the same questions Mythos is being asked. There is no source-code gate to defend. The threat surface is exposed by definition.

What can be defended is time. Whichever capability reaches the kernel first holds the window between vulnerability discovery and adversary weaponization. Federal and defense workloads run on Linux at enormous scale — embedded systems, network appliances, weapons platforms, intelligence infrastructure, federal cloud workloads built on Red Hat Enterprise Linux and its derivatives. Getting Mythos in front of the kernel before adversary AI capability matures fits the partner pattern exactly — it's the same strategic move, applied to a codebase that cannot be protected by closure.

Government involvement, in public

The role of the US government in Project Glasswing has been less hidden than overlooked. Press reporting has documented that National Cyber Director Sean Cairncross is leading a federal-officials group focused on critical-infrastructure vulnerabilities and AI exploitation risk, and Cairncross has publicly acknowledged White House meetings on Mythos's risks and benefits with industry. Dario Amodei met directly with the White House Chief of Staff and the Treasury Secretary. Reports also indicate the administration opposed plans to expand Glasswing membership by approximately seventy additional organizations — a posture that, if accurate, suggests the executive branch held de facto influence over the partner list, though "effective veto authority" is an inference from the reported pattern, not a stated arrangement. The Pentagon's cyber policy chief did publicly frame Mythos as a US industry innovation advantage, language that fits industrial policy more naturally than neutral cybersecurity research.

None of this is covert — it's reported in public sources. The surface framing of Glasswing as a private-sector defensive consortium plausibly understates how much of the program operates as US national-security industrial policy with government coordination. That reading is an inference from the publicly reported facts, not a position Anthropic has formally adopted. A narrower reading is also possible: the partner list may simply reflect where Anthropic had the strongest commercial, cloud, and security relationships at launch. The industrial-policy interpretation is the stronger strategic read for this brief, but it remains an inference, not a documented arrangement.

Sources for this subsection · CNBC on the Vance-Bessent call with tech CEOs (April 10, 2026) for the pre-launch CEO conference call · CNBC on the Amodei–Wiles White House meeting (April 17, 2026) for the post-launch engagement · Reuters on the European Commission–Anthropic engagement (May 4, 2026) for the Dombrovskis briefing and not-yet-available European bank access · CNBC on the EU–OpenAI Daybreak/Anthropic asymmetry (May 11, 2026) for the four-to-five EU–Anthropic meetings without a concrete access proposal · Reuters on the ASIC urgent-cybersecurity-action call (May 8, 2026) for the Australian regulator engagement. Treat publisher-attributed claims as press-reported under the framing block at the top of this section.

The governance gap

UPDATE · MAY 2026 — EVENTS DURING THE BRIEF'S REVISION WINDOW

Three developments during the brief's revision window bear directly on the central decoupling argument and on the regulator response. (1) On May 11, 2026, the head of the Bank of England's Prudential Regulation Authority, Sam Woods, told a UK Finance audience that it was "reasonable to expect quite significant disruption" to financial services from advanced AI models including Mythos and ChatGPT 5.5 Instant, citing the growing capability to identify vulnerabilities and the requirement for banks to patch them at speed; Woods called patching "the main driver of outages" in the financial system. (2) The IMF separately, on May 7, framed AI-powered cyber attacks as a financial-stability question, not merely an operational-risk one.

(3) Most importantly for the substantive argument of this brief: on May 11/12, Google's Threat Intelligence Group disclosed what it described as the first identified case of a threat actor using a zero-day exploit it believes was developed with AI — a 2FA bypass for a popular open-source admin tool, attributed to a cybercrime group, with the exploit code bearing several hallmarks of LLM authorship: docstrings, a hallucinated CVSS score, and textbook Pythonic structure. Google says the criminal threat actor planned to use the exploit in a mass-exploitation event and that its proactive counter-discovery may have prevented broad use; the disclosure is the first such case Google has identified, not necessarily the first in the world. This was not attributed to Mythos, and Google says it does not believe Gemini was used; the specific model and access path were not identified publicly. The narrow but important point: AI-assisted vulnerability discovery and exploit generation has moved from theory into live adversary workflows, and the narrow "no live adversary workflow exists yet" objection can be retired. The precondition behind the diffusion clock is no longer theoretical, even though Mythos-class capability has not been shown to be broadly available.

> Mythos changes the equation differently. Where currently accessible models are constrained primarily by the safety filtering built into general-purpose deployment — the brake Opus 4.7 makes explicit through its "differentially reduced" cyber capabilities — Mythos is controlled instead through restricted access, partner obligations, monitoring, and gated deployment. Anthropic has stated it is testing broader automated cyber safeguards first on less capable models. The publicly disclosed brakes on Mythos appear primarily organizational and contractual — restricted access, partner obligations, monitoring, and gated deployment — rather than a generally available product-safety layer, which matters because organizational brakes fail differently than capability-based ones: they fail in cases like the Bloomberg-reported contractor-channel access, not in cases like a public jailbreak. Both failure modes are now in evidence in the broader frontier-model category. Neither category is hypothetical.

For a program operating at this scale and stake, the publicly disclosed enforcement architecture is unusually thin.

WHAT ANTHROPIC STATES

The disclosed safeguards.

+ Partners subject to Anthropic's published Usage Policy.
+ Model provided exclusively for "defensive cybersecurity purposes."
+ Gated API access through Bedrock, Vertex AI, and Foundry with partner-only identifiers.
+ Monitoring probes deployed, though configured not to block exchanges during the research preview.
+ $100M in usage credits creates financial dependency that could in principle be revoked.
+ A Cyber Verification Program for legitimate security professionals, announced with Opus 4.7 on April 16, 2026. Eligible security professionals apply through an organizational application flow for adjusted treatment of legitimate dual-use security work; the program is free, and Anthropic's current help article notes platform-specific limitations including that Bedrock, Vertex AI, and ZDR accounts are not currently in scope.

WHAT IS NOT PUBLICLY SPECIFIED

The unaddressed gaps.

− Whether defense-only obligations are contractually binding with enforcement mechanisms, or aspirational.
− Whether Anthropic logs which codebases partners scan, and whether anomalous scanning patterns trigger review.
− What specific actions constitute partner misuse and what penalties attach to violations.
− Whether partner employees are individually vetted, and by whom.
− Whether independent auditing of partner usage occurs.
− What reporting requirements apply to partner-discovered vulnerabilities and on what timeline.

The trust model, in operational terms, is one layer deep. Anthropic trusts a defined set of organizations — the 12 launch partners plus over 40 additional ones. It does not appear to verify in any publicly disclosed way what those organizations actually do with the model on a continuous basis. The monitoring probes are explicitly configured to observe rather than block during the research preview. The Usage Policy applies but the enforcement architecture is not published.

This matters for a specific reason: the insider threat surface is large. Glasswing partners include financial institutions, cloud providers, and cybersecurity firms. Every one of these organizations has both defensive and offensive security teams. The same employees performing legitimate defensive vulnerability scanning could pivot to offensive research using the same authenticated access. Bloomberg reported unauthorized Mythos access through a third-party vendor environment; Anthropic said it was investigating the report and had not found evidence of broader impact to Anthropic systems. The governance question that report raises has not been publicly answered.

The broader observation is that Anthropic's own security posture during the same period has been imperfect. The CMS misconfiguration that exposed the Mythos draft, and the separate npm packaging error that leaked Claude Code source code, both happened within roughly a month of the Glasswing announcement. A program asking its partners to trust Anthropic with sensitive vulnerability data is being run by an organization that has demonstrated two recent failures of basic configuration hygiene. That is not a disqualifying observation, but it is a relevant one.

The CXO read

For organizations outside Project Glasswing, the practical implications are three.

First, your critical infrastructure vendors are likely partners — the major cloud platforms, your endpoint security vendor, your network equipment vendor, your laptop fleet manufacturer. That is defensively useful in the medium term, as their patching velocity should improve. It also means your security posture now depends on the patching velocity of organizations you do not control, scanning for vulnerabilities you cannot see, on a timeline you are not consulted on.

Second, your competitors with offensive security capabilities may also be partners, depending on industry. The official restriction is to defensive use. The actual enforcement of that restriction across thousands of employees with authenticated access is not publicly auditable. The competitive intelligence risk is non-zero, though it is probably lower than the patching benefit for most organizations.

Third, and most strategically: when comparable offensive capability diffuses outside Glasswing — a window this brief treats as six to eighteen months for planning purposes — the protected stack will be hardened first. Everyone else absorbs the impact of more widely available offensive AI cyber capability against a substrate that has not been pre-hardened in the same way. For non-partner enterprises, Glasswing access isn't on offer to most. The real question is whether to invest now in the remediation pipeline described in the previous section, on the assumption that the protective head start is not coming to you for free.

FOR NON-US-HEADQUARTERED ORGANIZATIONS

European, Asian, and Indian enterprises are downstream beneficiaries of Glasswing-driven patches to shared infrastructure, but are not visibly represented in the named launch-partner list. The composition of the additional 40+ organizations has not been fully disclosed, so this claim is about the visible Glasswing roster rather than the complete access footprint. Reuters has reported European Commission engagement with Anthropic and an Australian securities regulator urging urgent cybersecurity action — both of which read as outside-the-circle responses, not inside-the-circle preparation. Plan accordingly.

Glasswing is probably a reasonable program. It also rests on thinner public accountability than its stakes call for, and its framing as a private-sector defensive consortium understates how much of the architecture is industrial policy with government coordination. None of that makes the arrangement wrong. It does mean the public description of what Glasswing actually is has not kept up with the capability it's coordinating around.

Section 05 · What you can actually use today

The product you can buy, and what it leaves untouched.

Mythos itself is not generally available, and Anthropic has not announced a commercial release. The capability a CXO can actually procure this quarter sits in three distinct buckets, and the procurement conversation goes better when the buckets are kept separate. The category is no longer one vendor wide.

FRONTIER-LAB PRODUCTS *Direct AI-assisted code-security tools from the frontier labs*	Anthropic · Claude Security *(previously Claude Code Security)* OpenAI · Codex Security *(formerly Aardvark)* Google · Gemini Code Assist Enterprise + security extension Google · Big Sleep / CodeMender agents
AI-NATIVE REVIEW PLATFORMS *Independent vendors built around AI-driven code review*	CodeRabbit · Greptile · Qodo · Cursor BugBot

ESTABLISHED APPSEC + AI *Incumbent application-security vendors layering AI on existing pipelines*	Snyk Code	GitHub Advanced Security + Copilot Autofix
	Semgrep	SonarQube
	Checkmarx	

The right question is narrower than "should we adopt AI code security." It is which parts of the existing security stack these tools genuinely improve, and which parts they leave entirely untouched. Both halves of that answer matter for the procurement conversations that will happen over the next two quarters.

What the productized capability actually does

Claude Security, available in public beta for Claude Enterprise customers, performs three things noticeably better than the previous generation of static analysis tools. It reasons over data flow rather than pattern-matching for known signatures, which lets it identify bug classes that traditional SAST cannot articulate as rules. It generates suggested patches that a human reviewer can approve or reject, collapsing the latency between finding and proposed fix. And it produces higher signal-to-noise output when paired with a project-specific harness — Mozilla's Firefox 150 result, with its 271 fixes traceable to a single evaluation, is the proof of concept for what disciplined integration with existing test infrastructure can deliver.

OpenAI, Google, and Anthropic all now claim productized workflows around AI-assisted code reasoning, vulnerability validation, and patch suggestion. At procurement, the differences are model choice (Claude Opus, GPT-5.5, Gemini), depth of agentic harness, IDE and CI integration, and data-handling terms. The operational value of each tool will differ less by model label than by harness quality, false-positive controls, integration depth, and proof that fixes are verified in production; those dimensions remain product-specific and have not yet been independently benchmarked across frontier-lab security products. The independent Martian Code Review Bench, published in February, gave teams the first non-vendor comparison across the AI-native review platforms; no equivalent cross-lab benchmark yet exists for the frontier-lab security products themselves, which is itself a procurement signal.

What it does not do is replace any of the categories that performed real defensive work last quarter. Endpoint detection, identity governance, runtime application protection, web application firewalls, dynamic API testing, cloud posture management, and incident response platforms address problems that static code reasoning is not designed to solve. StackHawk's published critique of the category is technically correct on the substance: business-logic vulnerabilities, authorization failures, and session-handling bugs typically surface only when an application is running and an active testing process is probing actual behavior, not when source code is being read in isolation.

TOOL CATEGORY	WHAT IT DOES TODAY	AI TOOLING OVERLAP	WHAT REMAINS UNCOVERED
SAST	Finds code vulnerabilities from source through rules and pattern matching.	HIGH Reasoning over data flow outperforms signature matching on complex paths.	Project-specific harnessing; validation that findings are reachable in production.
SCA	Identifies vulnerable open-source dependencies via known CVE matching.	HIGH Reasons across dependency source and proposes upgrade or patch paths.	SBOM provenance, license compliance, supply-chain attestation.
DAST & API testing	Probes running applications and live APIs for behavior under attack.	PARTIAL Possible with scaffolding, but not a packaged feature of current AI tools.	Runtime state, auth flows, session handling, environment-specific behavior.
Pentesting	Authorized humans find and validate vulnerabilities in a scoped engagement.	MINIMAL Closed-source models carry offensive-use guardrails and lack the integrated attack-tool chain that engagements require. Mythos-class pentesting capability is not procurable today.	Everything substantive — judgment, exploit validation, rules of engagement, defensible reporting, and the offensive tooling itself.
EDR / XDR	Detects and responds to threats on endpoints and across telemetry.	MINIMAL Different problem domain; live telemetry and containment.	Everything. Discovery acceleration may make EDR more important, not less.
Identity / IAM	Governs who can access what, with what privileges, for how long.	MINIMAL Vulnerability discovery does not address lifecycle, session, or privilege.	Access governance, session controls, privilege management — entirely.
WAF / runtime protection	Blocks malicious traffic and exploit attempts against running applications.	MINIMAL Code reasoning is not a substitute for live enforcement.	The compensating control layer when patches are not yet deployed.

The jagged frontier

The most consequential finding about this category, for procurement purposes, did not come from Anthropic or its evaluators. It came from AISLE, an AI cybersecurity startup, which examined AI cybersecurity as a modular pipeline rather than a single model output. AISLE's conclusion — replicable on smaller and cheaper models for many narrower tasks — is that the durable advantage in AI-driven security tooling lies not in the model itself but in the system around it: scanning targets, validation logic, triage workflow, patch generation discipline, regression testing infrastructure, and maintainer trust. Once a relevant code path is isolated, much smaller models can recover similar results. The model is necessary. It is not where the moat lives.

> *The model is necessary. The moat lives in the system around it.*

The procurement implication is direct. A vendor whose only differentiation is "we have an AI in it" is offering a feature that will commoditize within twelve months as comparable models become broadly available through Bedrock, Vertex, Foundry, and successive open-weight releases. A vendor whose differentiation is the harness, the workflow integration, the patch verification pipeline, the false-positive controls, and the audit trail is offering something more durable. The right question in the next procurement conversation is not "which model do you use" but "which of these surrounding capabilities have you actually built, and what does the false-positive rate look like in production at our codebase scale."

For the next two quarters, the strategic decision is not whether to adopt an AI code security tool. Some version of one is arriving in every major enterprise security suite regardless of the choices made in the procurement conversation. The decision is which adjacent investments — dynamic testing, runtime controls, identity hardening, incident response capacity — to fund alongside it, because none of those categories are being displaced by the new tooling, and several of them become more important as discovery accelerates.

> THE PROCUREMENT TEST, THREE QUESTIONS
>
> *When a vendor demonstrates AI vulnerability discovery, ask three things. First, what is the false-positive rate when the tool runs unattended on a representative sample of your code, with the denominator disclosed. Second, what is the actual workflow from validated finding through deployed patch through verified-in-production, and how is that loop instrumented. Third, what compensating runtime controls do they recommend for the substantial majority of findings that will not be patched within the quarter. The quality of those answers separates a platform from a feature.*

Section 06 · Market view & vendor exposure

Repriced — and mispriced.

Two equity-market reactions to Mythos and its productized cousin have already happened, and a procurement officer can read both of them. The February selloff that followed the Claude Security launch (then named Claude Code Security) took JFrog, GitLab, Okta, CrowdStrike, Cloudflare, and Zscaler down together. The accidental Mythos leak on March 26 triggered a broad cybersecurity selloff, with several major vendors and the cybersecurity ETF falling materially within twenty-four hours; specific percentages varied across sources. Both moves contained real signal. Both also overshot, and the way they overshot is the most useful data point for anyone with a renewal in the next two quarters.

The signal both selloffs got right is narrow but real. AI reasoning over data flow, combined with patch suggestion, genuinely changes the unit economics of pattern-match application security tooling. Forrester's blunt framing — that the category that took "real damage" was the one perceived as dependent on pattern-matching SAST, SCA, and ASPM — is the most defensible part of the market reaction. For that slice of the security stack, the model repriced the labor.

Both selloffs got the scope wrong. The same downdraft hit endpoint detection vendors, identity governance vendors, network enforcement vendors, and cloud-security posture vendors — categories that Claude Security and Mythos do not replace and, in several cases, become more necessary in a world of accelerated discovery. A Reuters-quoted analyst stated the technical point plainly: Claude Security does not handle live intrusion detection, does not stop attacks in progress, and does not analyze compiled production components. The same architectural ceiling applies to Codex Security, Gemini Code Assist with its security extension, and every AI-native code review platform in the category — none of them are runtime products, and none of them claim to be. The repricing of AppSec economics is not the same event as a repricing of the security stack.

MOST EXPOSED *AppSec discovery economics — genuine repricing*	Traditional SAST · SCA · ASPM dashboards · Commodity pentest discovery
PARTIALLY EXPOSED *Overlap with code reasoning, but retain independent value*	DAST & runtime testing · Fuzzing · Bug bounty platforms · CNAPP / cloud posture · External ASM

LESS EXPOSED

Often more important as discovery accelerates

EDR / XDR | Identity — IAM / IGA / PAM | WAF & API protection | SIEM / SOC / MDR | SOAR / IR platforms

A second tell: several of the vendors that sold off most sharply on the Mythos leak are themselves Glasswing partners. CrowdStrike and Palo Alto Networks, both inside the consortium with privileged early access to the very capability the market was reacting to, dropped alongside vendors that have no such access. That doesn't read as efficient pricing of disruption risk. It reads as panic in a category where the sell-side analyst community could not yet distinguish the AppSec story from the rest of the security stack. Whether the market walks those positions back will be visible in two-quarter and one-year price action, and how the named vendors describe the AI-displacement narrative in their next earnings cycles is the test — this brief treats the panic as the more reliable signal than any walk-back narrative built on top of it.

Where the durable value lives

The likely winners over the next six to eighteen months are the platforms that own the post-discovery workflow rather than the discovery step itself. The discovery layer is commoditizing — that part of the market reaction will eventually be priced correctly. The work that does not commoditize is what surrounds discovery: asset inventory and reachability analysis, exploitability validation, patch generation paired with regression testing, deployment automation, runtime compensating controls, and the governance and audit trails that regulators and insurers will increasingly require. A vendor that is strong across that stack and incorporates AI into its discovery layer holds the strongest hand. A vendor whose only differentiation is discovery — AI-assisted or not — is the most exposed.

The procurement signal

For a CXO with security renewals in the next two quarters, the asymmetry is exploitable but narrow. Pricing pressure is genuinely available on pure-play AppSec, SAST, SCA, and ASPM contracts, especially on multi-year commitments where the vendor's value proposition has narrowed. The harder negotiation is on the broader platforms — EDR, identity, WAF, SIEM — which are arguably more important now than before, and which the vendors themselves know it. The procurement task is to renegotiate AppSec spend without conceding leverage on the categories that should sustain or expand.

> WHAT TO SAY IN THE NEXT RENEWAL
>
> *For AppSec, SAST, and SCA renewals, ask the vendor to disclose how they intend to compete when comparable AI discovery capability is available through Bedrock, Vertex, or Foundry at a fraction of the seat price, and to demonstrate which workflow integrations justify their premium. For EDR, identity, WAF, and SIEM renewals, do not let equity-market noise become a vendor-side negotiation talking point. Those tools are doing different work, and the work is becoming more important.*

Section 07 · Actions by horizon

What to do, and when.

An action plan that holds up starts with measurement, not procurement. Nothing in the thirty-day list below is technologically difficult. All of it is organizationally difficult — inventory, baseline measurement, and governance of AI tools that engineering teams are already using without security oversight — and the difficulty is exactly the point.

A note on sequence. These three lists are designed to compound. Skipping the thirty-day measurement work makes the sixty-day pilot harder to evaluate, because there is no baseline against which to read the result. Skipping the sixty-day work makes the four-month reorganization conjectural rather than instrumented. The ordering matters more than the specific tools chosen along the way.

A NAMED FRAMEWORK

The Remediation Throughput Maturity Model

The horizon cards below describe a sequence. The four levels of that sequence have explicit criteria, and an organization can locate itself on the model from observable facts about its current operating cadence. Movement is from Level 1 to Level 4. Most enterprises start at Level 1; the four-month plan that follows is a Level 1 → Level 2 jump.

LEVEL 1

Discovery-led

Investment concentrated in scanning, threat intelligence, and EDR. MTTR for critical OSS dependencies measured in weeks to months and not systematically tracked. Patch cadence calibrated to the prior threat-landscape equilibrium.

LEVEL 2

Triage-capable

Mean time from validated finding to verified production fix is measurable and reported. Patch deployment runs on a known cadence with an emergency-override path for critical security-only changes. Compensating controls exist where patching is delayed, though not yet automated.

LEVEL 3

Throughput-disciplined

MTTR is a board-reported metric. Patch deployment is automated for non-breaking changes. Regression-test suites for security patches run continuously. The organization can absorb a quarterly surge in disclosed findings without backlog growth.

LEVEL 4

Continuous-flow

Discovery, validation, patching, verification, and compensating-control activation run as one connected pipeline. Failure modes are explicit and have automated fallbacks. The organization can absorb machine-pace discovery without backlog growth, and can demonstrate loop closure inside an audit window.

30 *days*

MEASURE & MAP

Make the gap visible inside your own organization.

01 **Baseline the estate.** Inventory the open-source and operating-system components actually running in production, then rate each on blast radius and severity-if-compromised. The output is a ranked shortlist of ten to twenty candidates that justify immediate attention — the input to every later horizon. Most organizations have never produced this list.

02 **Measure the real number.** Calculate mean time from validated finding to verified-in-production fix — not finding-to-ticket-closed, not finding-to-patch-released. Verify that SAST, SCA, DAST, secret-scanning, and gated review are actually deployed on the critical tier. Most dashboards overstate coverage; the baseline is what counts.

03 **Map dependencies and tools.** Identify which critical-infrastructure vendors sit inside Project Glasswing, audit the AI tools engineering teams are already using to read and generate code, and brief the board on a single question: can fixes close faster than adversaries can weaponize the same class of capability within six to eighteen months?

60 *days*

PILOT & CONTAIN

Move from measurement into instrumented practice.

01 **Pilot AI-assisted review.** Run Claude Security or a frontier-lab equivalent — or an AI-native platform layered onto the existing SAST stack — on one critical codebase, with a disclosed false-positive denominator. Stand up AI-assisted triage lanes for three-to-ten times the prior finding volume, and require reproducible test cases before any AI-generated finding enters the patch pipeline.

02 **Collapse vendor-patch absorption.** Move from days to hours on the critical tier — the companion playbook's P0 remediation SLA is 24–72 hours or an approved compensating control. Automated dependency resolution, pre-staged regression suites, unattended deployment for security-only changes, and a patch-approval channel separate from the feature-release change board. Scope a substitution or accelerated-patching path on the top one or two components — accelerated patching plus hardening for most Immediate components; substitution scoped only where independently validated.

03 **Contain blast radius.** Strengthen runtime compensating controls (WAF, segmentation, EDR coverage, identity hardening) for the majority of findings that will not be patched in this window. Make the line between critical and non-critical tiers architectural — enforced segmentation, non-shared trust, distinct credential stores. Tabletop a Mythos-class adversary scenario and renegotiate AppSec/SAST/SCA contracts falling due before the next budget cycle locks the prior pricing in.

120 *days*

INDUSTRIALIZE & MEASURE

Reorganize the security function around the new bottleneck.

01 **Industrialize the pipeline.** CI/CD-integrated patching, continuous regression testing, and an evidence-backed verification loop where P0/P1 closure requires proof appropriate to finding type. Stand up the remediation council, pre-CVE intake, and AI-discovery governance — the operating fabric that lets the program absorb machine-pace finding volume without backlog growth.

02 **Scale substitution selectively.** Convert the sixty-day pilots into a multi-year dependency-substitution program scoped tightly to the top tier. Most open-source dependencies will and should stay; replacement applies only where blast radius does not survive a worst-case discovery event and the alternative has been independently validated, not just procured.

03 **Make the metric board-grade.** Reassess the security vendor portfolio against post-discovery workflow ownership — discovery-only differentiation is the most exposed and replaceable. Promote mean-time-to-deployed-fix to an executive metric reported at the same cadence as availability or revenue. What is measured at that level gets resourced; what is not, does not.

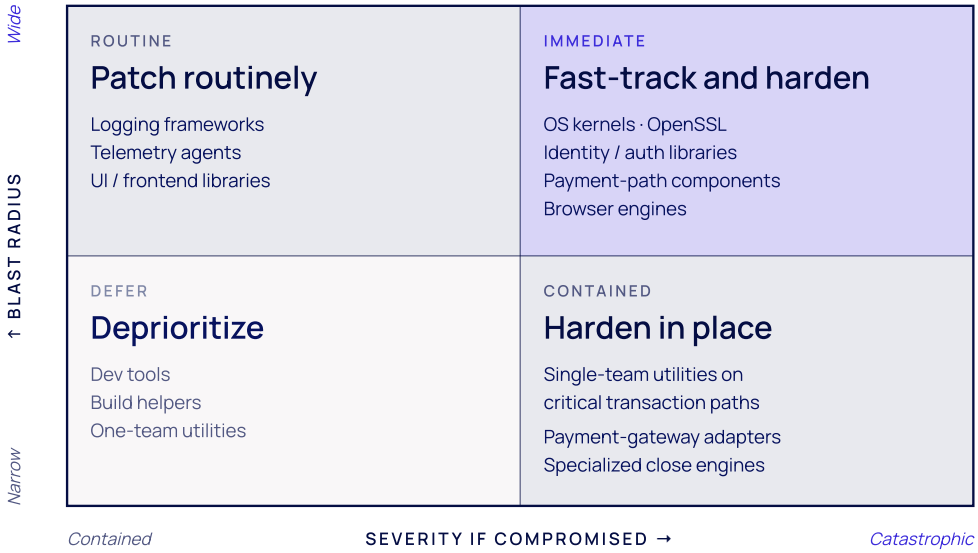
FOR THE DETAILED EXECUTION PLAN

The summaries above are the executive horizon. The full operating plan — weekly milestones, owners, decision gates, risk-tier rubrics, SLA models, routing rules, RACI, ticket payloads, verification rulebook, program cadence, and Day 30 / Day 60 / Day 120 exit gates — is maintained as a companion playbook (the *AI-Scale Vulnerability Management 30/60/120-Day Execution Plan* and the *Remediation Throughput Playbook*).

To request the detailed action plan, the workstream-level RACI, or a tailored walkthrough for your environment, contact **marketing@hexaware.com**.

Remediation candidate selection rubric

Author analysis — not a measured dataset. Output of the thirty-day inventory and rating exercise. Rate each open-source or operating-system component on blast radius and severity-if-compromised; the response category follows from the placement.



How to use this · Place each open-source or operating-system component from the thirty-day inventory into a quadrant. The placement determines the response category and the verb that governs the next sixty days. Components in the top-right quadrant are the candidates for accelerated patching, hardening, and — where technically realistic — substitution piloting. Components in the lower-left are the queue items that should not consume incident-response capacity. Examples shown are illustrative; the actual placement depends on each organization's architecture, customer surface, and revenue dependencies.

THE ONE-LINE TEST FOR ANY ITEM ON THIS LIST

If an action item does not reduce the time between a validated finding and a verified fix in production, or reduce the number of unpatched findings exposed to an adversary with comparable discovery capability, it is not on this list for a reason that survives executive scrutiny. The bottleneck has moved. The budget, the org chart, and the operating cadence have not yet. Closing that gap is the work.

Section 08 · Honest synthesis

The shift, not the tool.

> *Can your organization absorb, validate, prioritize, patch, test, and deploy fixes faster than adversaries can weaponize the same class of capability?*

Today, in almost every enterprise, the answer is no. The window to change that answer is finite — this brief treats six to eighteen months as a planning assumption, informed by public expert commentary and frontier-lag analysis, including Alex Stamos at RSA 2026. After that window closes, comparable offensive capability may be accessible to adversaries against substrates that have not been pre-hardened. The protective head start that Glasswing partners enjoy does not extend automatically to the rest of the economy.

Buying a tool doesn't close this gap. No product on the market today does, and no product arriving over the next twelve months will either. What works is an industrial reorganization of the security function around a new bottleneck. Discovery is commoditizing. Everything downstream of it — validation, prioritization, patching, regression testing, deployment, runtime compensating controls, governance, and proof of risk reduction — is becoming the premium tier. The reorganization is the work of the next four months. The tool selection is incidental.

CXOs who reorganize on remediation throughput, internal AI-tool governance, and runtime compensating controls during the diffusion window will be measurably ahead when comparable capability becomes available outside Glasswing. Those who treat the question as a vendor selection problem will be behind. The decision is organizational, and it has to be made this year.

An uncomfortable corollary of the partner concentration

The conventional read of Project Glasswing is that giving defensive AI capability to a small set of high-leverage organizations is net positive for cybersecurity: the substrate gets hardened first, and downstream consumers inherit the benefit. That read is partially correct. It is also incomplete.

The publicly named launch partners are not chosen at random from the population of organizations that defensive AI capability would most benefit. They are chosen, by both their selection and their composition, from the population already best resourced to defend itself: hyperscalers, frontier security vendors, primary clearing banks, a kernel foundation. The composition of the undisclosed 40+ additional organizations is not public, so this claim is about the visible launch-partner list and press-reported additions, not the full Glasswing footprint. The marginal defensive benefit of giving an AI cyber tool to AWS, CrowdStrike, or JPMorgan is real, but smaller than the marginal benefit of giving the same tool to a regional hospital network, a mid-tier SaaS provider, or a sovereign-cloud operator with thinner security staffing. The diffusion clock then runs the other way: the offensive

side of the same capability becomes broadly accessible on a months-not-years timeline, and the long-tail enterprises that did not get the defensive head start absorb the impact first.

That implies a reading worth making explicit, even though it cuts against the dominant framing of the program. Glasswing's defensive concentration may leave the aggregate non-partner stack *relatively* less safe, not more, because the marginal defense went where the marginal benefit was smallest. The substrate-hardening counter-argument is real — Linux Foundation patches do propagate downstream, AWS hardening does help its customers, browser-engine fixes do reach every user of the browser. But that counter depends on the substrate-layer effects being larger than the missing direct-defense effects on the unprotected middle tier. The brief takes this contrarian read seriously rather than dismissing it, and treats it as a hypothesis the next eighteen months will test.

One caveat is important enough to state directly. The claim is comparative, not absolute. It does not say Glasswing makes non-partners less safe in absolute terms; the likely benefit of upstream hardening is real, and a non-partner enterprise running on Glasswing-hardened substrate is almost certainly better off than it would be on un-hardened substrate. The claim is that scarce defensive capability may have been allocated to organizations with lower marginal need than the unprotected middle tier — meaning the aggregate distribution of defense is less efficient than a uniformly-distributed alternative would have been. The reader should hold those two things separate.

What this brief is willing to be checked on

A brief that hedges every claim is hard to falsify. The two predictions below are testable from publicly observable data. The brief stakes its analytic credibility on them.

PREDICTION 1 · DIFFUSION SIDE

By the end of Q1 2027, at least one openly available or broadly accessible non-Western model will demonstrate Mythos-like capability on public cyber-evaluation tasks covering at least two of these target classes: browser engines, OS kernels, network services, or major open-source infrastructure libraries. "Mythos-like" here means not just flagging candidate issues, but producing reproducible proof-of-concept exploits with comparable signal-to-noise on at least one disclosed bug class per target. If this does not happen, the 6–18 month diffusion-window framing in this brief should be revised upward.

PREDICTION 2 · REMEDIATION SIDE

By the end of October 2026, the credible public data streams on enterprise remediation cadence — the Cyentia Institute's *State of Exploit Prediction*, Veracode's *State of Software Security*, CISA KEV-derived remediation timing, and comparable insurer or large-vendor telemetry are the streams this brief expects to look at, though others may emerge — will still show critical-vulnerability remediation for Immediate-tier OSS dependencies measured in weeks rather than days. If credible audited enterprise data instead shows median verified-production remediation falling below seven days for Immediate-tier OSS dependencies, the central decoupling argument in Section 03 should be revised.

What would change my read

Three observable developments would push this brief to revise its central claims in non-trivial ways. They are listed for the reader who wants to know what evidence would falsify the analysis, not confirm it.

→ *Major-enterprise MTTR data tracking discovery output.* If a Fortune 100 organization publishes MTTR data — under audit, not marketing — showing that its remediation cadence has matched or exceeded the discovery-output curve from Project Glasswing across its critical OSS dependencies, the discovery/remediation decoupling thesis weakens substantially. The brief currently assumes the gap is widening for almost every enterprise. A single credible counterexample at scale, properly documented, would force a rewrite of Section 03.

→ *Anthropic publishes a follow-on red-team evaluation reconciling the "thousands" figure.* If Anthropic publishes a deeper analysis showing that Mythos's candidate vulnerabilities concentrate heavily in non-remotely-exploitable categories under realistic mitigations, the "thousands of zero-days" framing should be dialed back further than this brief currently dials it. The 89%-severity-agreement validation is necessary for that question but not sufficient.

→ *Diffusion fails on schedule.* If no openly available or broadly accessible model reaches Mythos-class capability on common target classes by the end of Q1 2027, the lower-bound diffusion assumption weakens. If no such model appears by October 2027, the full 6–18 month planning window was wrong in the wrong direction. The right fix is not to celebrate the longer head start; it is to ask why the assumed diffusion model failed, and whether the asymmetry is structural rather than temporal.

Appendix · References & further reading

Where this came from, and where to read next.

Coverage of Mythos has come from a wide mix of sources with very different commercial interests, and the single most useful filter for reading future coverage is who benefits from a particular conclusion being widely believed. The pages below sort the major sources used in this brief by that filter, define the acronyms used throughout, and flag three voices worth reading directly.

How to read the sources

The split is not a quality ranking. Vendor and analyst sources contain useful and sometimes essential information — Anthropic's own technical report is the most detailed account of Mythos's capabilities, and Forrester's market framing has been correct about which categories are most exposed. The point of separation is that independent and commercially interested sources should be weighted differently when their conclusions diverge, and those divergences are the most informative data.

INDEPENDENT / DISINTERESTED

Weight conclusions heavily.

UK AI Security Institute (AISI). Government evaluator. Published methodology, controlled environments, named limitations including the absence of active defenders in their test ranges.

Mozilla Corporation. Operator with disclosed harness design, severity breakdown, and reproducibility methodology. The 271 Firefox 150 number arrives with the operational detail that supports it.

AISLE. Independent research that replicated portions of Mythos's analysis on smaller, cheaper models. The "jagged frontier" finding is the most consequential skeptical contribution to date.

Bruce Schneier. Multi-decade track record on dual-use technology and capability proliferation. Independent of every commercial stakeholder in the conversation.

Daniel Stenberg. Maintainer of curl since 1998. Posted the first inside-the-program review of a Mythos scan, including hard numbers on what survived independent verification. Independent of every commercial party in the conversation. Splits the marketing from the category clearly — "not particularly dangerous" on the hype, "significantly better than traditional analyzers" on the technology underneath it.

Cal Newport. Georgetown computer scientist who reads the announcement through the lens of prior AI benchmark-tuning cycles rather than capability hype.

VENDOR OR COMMERCIALLY INTERESTED

Useful — but interrogate the framing.

Anthropic. Model maker and program lead. Technical detail in the Frontier Red Team report and risk assessment is essential reading; the framing reflects the company's own commercial and policy positioning.

Forrester. Analyst firm. Strong on market reaction and category-level vendor exposure; the "SaaS-pocalypse" framing is commercially attention-driven even when directionally correct.

StackHawk. Direct competitor in dynamic application security testing. Their critique of static-only code reasoning is technically valid; the commercial motivation neither establishes nor invalidates it.

ArmorCode, ZeroFox, Cobalt, Strobes. Vendor analyses written to map their own products onto the news. Read for what they describe, not for the conclusions they draw.

> **News reporting (Reuters, Fortune, BBC, CRN, Politico).** *Trust what reporters observe directly — stock movements, government meetings, partner lists. Treat what they quote others saying as second-hand testimony to be verified against primary sources.*

Glossary

Terms used in this brief, in the order a CXO is most likely to encounter them.

Project Glasswing	Anthropic-led consortium giving its 12 launch partners and over 40 additional organizations restricted access to Claude Mythos Preview for defensive security work. Backed by $100M in usage credits and subject to Anthropic's Usage Policy.
Mythos Preview	Anthropic's restricted frontier model with significant cyber and agentic capabilities. Not generally available. Distinct from Glasswing (the program) and Claude Security (the productized capability).
Claude Security	Anthropic's productized AI-assisted code security workflow for Claude Enterprise customers. Previously known as Claude Code Security. Public beta, accessed via the Claude.ai sidebar or claude.ai/security. Reasons over code to validate findings and suggest patches for human approval, with patch handoff into Claude Code on the Web.
SAST `Static Application Security Testing`	Tooling that reads source code to identify vulnerabilities without executing the application. The category most directly repriced by AI reasoning over data flow.
DAST `Dynamic Application Security Testing`	Tooling that tests a running application from outside — what an attacker would actually probe. Runtime state, auth flows, and environment-specific behavior surface here, not in static analysis.
SCA `Software Composition Analysis`	Identification of vulnerable open-source dependencies through CVE matching and SBOM analysis. Adjacent to SAST in exposure to AI disruption.
ASPM `Application Security Posture Management`	Dashboards and platforms that aggregate findings across multiple security tools. Value depends on the underlying scanners; commoditizes alongside them.

EDR / XDR Endpoint & Extended Detection and Response	Live telemetry and response on endpoints (EDR), extended across network, cloud, and identity signals (XDR). Not replaced by Mythos and arguably more important when discovery accelerates.
CNAPP Cloud-Native Application Protection Platform	Integrated cloud security covering misconfiguration, identity graphing, runtime exposure, and policy enforcement across cloud workloads.
SIEM / SOC / MDR	Security Information & Event Management (the platform), Security Operations Center (the team and discipline), Managed Detection and Response (the outsourced version). The monitoring and response layer that absorbs the queue when patches are delayed.
WAF Web Application Firewall	Network-edge enforcement that blocks malicious requests against web applications. A compensating control when underlying code remains unpatched.
IAM / IGA / PAM	Identity Access Management, Identity Governance & Administration, Privileged Access Management. The lifecycle, governance, and elevation controls for who can do what — entirely outside Mythos's discovery scope.
CVE Common Vulnerabilities and Exposures	The public identifier and registry for disclosed vulnerabilities. The throughput of this system is one of the bottlenecks under pressure from machine-scale discovery.
Zero-day	A vulnerability not yet known to the vendor or the public. The "zero days" refers to the warning time defenders have before potential exploitation.
Harness	In security AI, the scaffolding around a model — test infrastructure, dynamic case generation, validation logic, and triage workflow — that converts raw model capability into reproducible findings. Mozilla's harness is the canonical public example.
MTTR Mean Time to Remediate	The average time from a validated vulnerability finding to a deployed and verified fix in production. The single operational metric most worth driving down inside the diffusion window.
Evidence tier as used in this brief	In Section 02, "evidence tier" sorts public claims about Mythos by source quality: **Tier 1** independently validated, **Tier 2** vendor methodology, **Tier 3** ceiling / what could not be done. This is a frame for reading external claims, not a ticket-level taxonomy. The companion *Remediation Throughput Playbook* uses two separate, distinct schemes inside the operating model: an evidence-tier

	scheme **E1–E4** for ticket-level claim reliability (scanner → reproduced → independently confirmed → exploit demonstrated), and a risk-tier scheme **P0–P3** for remediation priority and SLA. The three schemes serve different purposes and should not be conflated.
Risk tier `P0 · P1 · P2 · P3`	The remediation-priority scheme used in the companion *Remediation Throughput Playbook*. Drives owner-assignment and fix SLAs: P0 (1 business-day owner SLA, 24–72 hour fix or compensating control), P1 (2 business-day owner SLA, 14 calendar-day fix), with P2 and P3 on lengthened windows. Where this brief refers to "critical-tier" or "Immediate" components, the operating equivalent in the playbook is P0/P1.
ASL `AI Safety Level`	Anthropic's internal classification framework for model capabilities and the safeguards required at each level. Anthropic has described Mythos as having heightened cyber capability that warrants restricted deployment; the specific ASL designation has not been re-confirmed publicly in the materials this brief draws on.

Three voices worth reading directly

Coverage of Mythos varies widely in quality. The three pieces below are worth reading in full rather than in summary, because each carries a distinct skeptical lens that the rest of the press repeats without crediting.

Bruce Schneier

Schneier on Security · "Mythos and Cybersecurity" · April 13, 2026

Called Mythos *"very much a PR play by Anthropic — and it worked."* The skepticism is two-layered. Methodologically, he flags missing false-positive denominators and undisclosed reviewer agreement on what counts as a finding. Structurally, he points to the concentration of frontier offensive capability inside fewer than a hundred organizations with no public accountability layer. The most respected name on the skeptical side; the single piece to read if you read only one.

AISLE

AISLE · "AI Cybersecurity After Mythos: The Jagged Frontier" · April 2026

Demonstrated empirically that smaller, cheaper models can recover much of Mythos's analysis when given targeting context. A 3.6B-parameter model identified the FreeBSD buffer overflow; a 5.1B-active model recovered the OpenBSD SACK chain. Their conclusion — *the moat is the system around the model, not the model itself* — is the most consequential

finding for procurement and competitive analysis. Read with their own caveat in mind: they gave the smaller models hints and scoped context, so the comparison is an upper bound rather than head-to-head.

Cal Newport

calnewport.com · "Is Claude Mythos 'Terrifying' or Just Hype?"

Frames Mythos as plausibly *"a version of Opus 4.6 tuned to perform better on a handful of benchmarks."* The strongest version of the "benchmark gaming with a press strategy" argument, from a Georgetown computer scientist who has been measured about previous rounds of AI capability inflation. Worth reading specifically for the version of skepticism that takes the announcement at its own evidentiary word and finds it lacking.

Full source index

A NOTE ON SOURCES & ATTRIBUTION

This analysis is based on public sources and author interpretation. References to Anthropic, Google, Mozilla, OpenAI, and other named third parties are for analysis only and do not imply endorsement, affiliation, or partnership beyond the publicly cited facts. Stock-price movements and analyst positioning are reported as market signal and are not investment advice. Where sources conflict, the brief explains the conflict and treats the most independent source as the reference point.

Sources are grouped by the evidence category they live in. The same labeling is used in Section 04's framing block. Direct links are provided where a stable public URL exists; where reporting is paywalled, behind a subscription, or attributed to a publisher without a single canonical URL, the entry names the publisher and date only.

COMPANY-CONFIRMED · ANTHROPIC PRIMARY SOURCES

- Anthropic, Project Glasswing announcement (April 2026). Launch partners, extended-organization count, $100M credits, $4M open-source donations, defensive-use framing, per-token pricing.
- Anthropic Frontier Red Team, Claude Mythos Preview (April 2026). 198-case manual review, 89% exact severity agreement, 98% within one level, FFmpeg re-analysis, Linux exploitability nuance, >99% unpatched figure.
- Anthropic, Claude Security is now in public beta. Product naming change (Claude Code Security → Claude Security), Enterprise availability.
- Anthropic, Usage Policy. Usage-policy basis for defensive-use restrictions referenced in the brief.
- Anthropic, Introducing Claude Opus 4.7 (April 16, 2026). Cyber Verification Program announcement for legitimate security professionals.

INDEPENDENTLY VALIDATED · EXTERNAL TECHNICAL EVALUATIONS

- UK AI Security Institute, Our evaluation of Claude Mythos Preview's cyber capabilities (April 2026). 32-step attack evaluation, 3-of-10 end-to-end completion, 22/32 step average, 73% CTF result, active-defender caveat.
- Mozilla, The zero-days are numbered (May 2026). 271 Firefox 150 fixes, 423 April security-bug context, rollup-CVE explanation.
- Mozilla Hacks, Behind the Scenes Hardening Firefox with Claude Mythos Preview (May 2026). Harness architecture, dynamic test-case generation, triage workflow, release integration.
- Firefox 150 release notes (21 April 2026). Release date and security-fixes entry; specific 271 / 423 counts are sourced to the Mozilla blog and Mozilla Hacks entries above.

PRIMARY TECHNICAL DISCLOSURES · THREAT-INTELLIGENCE AND VENDOR TECHNICAL REPORTS

- Google Cloud / Google Threat Intelligence Group, GTIG AI Threat Tracker: Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access (May 12, 2026). Primary source for the first GTIG-identified case of a zero-day exploit that GTIG believes was developed with AI; describes the disrupted mass-exploitation plan, the 2FA-bypass technical details, and the LLM-authorship signals. Secondary coverage at The Hacker News.

POLICY, REGULATOR, AND FINANCIAL-STABILITY SOURCES

- IMF, Financial Stability Risks Mount as Artificial Intelligence Fuels Cyberattacks (May 7, 2026). Financial-stability framing for AI-enabled cyber risk; cited in the Section 04 May update.
- Reuters, Britain's bank regulator expects 'quite significant disruption' from latest AI models (May 11, 2026). Sam Woods / Bank of England PRA warning, cited in the Section 04 May update.
- Reuters, European Commission is in contact with Anthropic on Mythos, Dombrovskis says (May 4, 2026; Reuters wire syndicated). EU Commission has met with Anthropic and is assessing implications; Mythos not yet available to European banks.
- Reuters, Australia regulator calls for urgent cybersecurity action to counter Mythos (May 8, 2026; Reuters wire syndicated). ASIC commissioner Simone Constant: "the clock is at a minute to midnight."

PRESS-REPORTED · NEWS ORGANIZATIONS

- TechCrunch, Unauthorized group has gained access to Anthropic's exclusive cyber tool Mythos, report claims (April 2026). Anthropic statement that it is investigating.
- Bloomberg, Bessent, Powell Summon Bank CEOs to Urgent Meeting Over Anthropic's New AI Model (April 10, 2026; paywalled). Treasury-convened April 7 meeting at Treasury headquarters; Wall Street CEOs warned to take Mythos seriously and deploy it for vulnerability detection.
- Bloomberg, original reporting on the Discord-group unauthorized-access report (April 21, 2026; paywalled). Bloomberg corroborated the access with screenshots and a live demonstration; TechCrunch carries the publicly accessible summary.
- CyberScoop, Security leaders say the next two years are going to be 'insane' (RSA 2026). Alex Stamos “every nineteen-year-old in St. Petersburg” framing; months-not-years posture.
- CNBC, Cybersecurity stocks fall on report Anthropic is testing a powerful new model (March 27, 2026). CrowdStrike -7%, Palo Alto Networks -6%, Zscaler -4.5% on the Fortune leak day.
- CNBC, White House summons tech CEOs over Mythos cyber threat (April 10, 2026). Pre-launch Vance-Bessent CEO conference call.

- CNBC, Amodei meets with White House officials on Mythos (April 17, 2026). Post-launch Amodei–Wiles engagement.
- CNBC, EU–OpenAI Daybreak access contrasted with Anthropic Mythos talks (May 11, 2026). EU–Anthropic meeting cadence without a concrete access proposal.
- Euronews, Hackers breach Anthropic's "too dangerous to release" Mythos AI model, report (April 22, 2026). Parallel European-perspective coverage of the Discord-channel access and Goldman/Citi/BoA/Morgan Stanley testing.
- Euronews, What is Anthropic's Mythos? The leaked AI model that poses 'unprecedented' cybersecurity risks (March 30, 2026). Post-leak Mythos primer.
- Fortune, Exclusive: Anthropic 'Mythos' AI model representing 'step change' in power revealed in data leak (March 26, 2026). The original CMS-misconfiguration leak that exposed the draft Mythos blog post; the trigger for the March 27 cybersecurity selloff.
- The Hill, Anthropic's Mythos puts DC, Wall Street on high alert (April 2026). National Cyber Director Sean Cairncross leading a federal-officials group on critical-infrastructure vulnerabilities and AI exploitation risk; cited in Section 04.
- CyberScoop, Executive orders likely ahead in next steps for national cyber strategy (April 2026). Cairncross's public acknowledgment of White House meetings on Mythos's risks and benefits.
- Politico, BBC, and CRN coverage on stock-market reactions, partner announcements, and policy context. These outlets are referenced for general context only; no specific claim in this brief depends on them as primary evidence. Where a specific quote or figure is load-bearing, the primary publisher is linked directly above.

COMPARISON-SET PRODUCT REFERENCES · SECTION 05 CONTEXT

- OpenAI, Daybreak | OpenAI for cybersecurity (May 2026). Codex-Security-plus-GPT-5.5 cyber-defense initiative; partner list and three-tier model access framework.
- OpenAI, Codex Security: now in research preview. Application-security agent capable of detecting, validating, and patching vulnerabilities; rolling out in research preview to ChatGPT Pro, Enterprise, Business, and Edu customers.
- OpenAI, Introducing Aardvark. Establishes that Aardvark became Codex Security.
- Google DeepMind, Introducing CodeMender. Big Sleep / OSS-Fuzz / CodeMender security-agent framing referenced in Section 05.
- Google, Gemini Code Assist Enterprise product documentation. Security extension feature set referenced in Section 05.
- Martian, Code Review Bench — public GitHub repository. The cross-platform AI-native code-review benchmark cited in Section 05.

ANALYST INTERPRETATION · SECURITY RESEARCHERS AND POLICY ANALYSIS

- Bruce Schneier, On Anthropic's Mythos Preview and Project Glasswing, Schneier on Security (April 13, 2026). "Very much a PR play by Anthropic — and it worked"; missing false-positive denominators; concentration of frontier offensive capability inside a small group of organizations.
- AISLE, AI Cybersecurity After Mythos: The Jagged Frontier (April 2026). Empirical replication using smaller open-weight models against the same vulnerability locations; “moat is the system around the model” framing.
- Cal Newport, Is Claude Mythos "Terrifying" or Just Hype?, calnewport.com (April 2026). Benchmark-tuning skepticism, Georgetown computer-science perspective.

- Institute for AI Policy and Strategy (IAPS), Mythos and the Evolving Cyber Landscape: Implications and Policy Priorities (April 2026). Disclosure-to-exploitation compression data: 3,529 CVE-exploit pairs from CISA KEV, VulnCheck KEV, and XDB.
- Forrester, Project Glasswing Shows That AI Will Break The Vulnerability Management Playbook. The "playbook breaking" framing the brief references is from this post.
- Forrester, Project Glasswing: The 10 Consequences Nobody's Writing About Yet. Second- and third-order effects across security tooling, vulnerability management, insurance, and regulation; complements the playbook post.
- StackHawk, general technical materials on the limits of static-only code reasoning relative to dynamic application security testing: Modern DAST positioning and What to Know About DAST. The brief uses the StackHawk position as a representative DAST-perspective argument that runtime-only vulnerability classes (authorization failures, business-logic flaws, session handling) are not addressed by static code reasoning; the position is theirs, not a specific Mythos-response post.
- ArmorCode, ZeroFox, Cobalt, Strobes, and other vendor-side analyses. Read as positioning material, not as independent evidence.

AUTHOR INFERENCE · READINGS, NOT FACTS

- The framing of Project Glasswing as US national-security industrial policy with government coordination is the author's reading of the publicly reported facts (partner geography, White House and Treasury engagement, Pentagon language), not a position Anthropic has formally adopted.
- The description of administration influence over the partner-list expansion as a “de facto influence” or de facto veto is an inference from the reported pattern, not a stated arrangement.
- The 6–18 month diffusion-window cell on the diffusion clock is a planning assumption used in this brief, not a forecast endorsed by any cited source.
- The remediation-rubric quadrants and the “mean time from validated finding to verified production fix” metric are author analysis offered as decision aids, not measured industry benchmarks.

HEXAWARE

About Hexaware

Hexaware is a global technology and business process services company. Our 28,400 Hexawarians wake up every day with a singular purpose; to create smiles through great people and technology. With this purpose gaining momentum, we are well on our way to realizing our vision of being the most loved digital transformation partner in the world. We also seek to protect the planet and build a better tomorrow for our customers, employees, partners, investors, and the communities in which we operate.

With 45+ offices in 19 countries, we empower enterprises worldwide to realize digital transformation at scale and speed by partnering with them to build, transform, run, and optimize their technology and business processes.

Learn more about Hexaware at www.hexaware.com