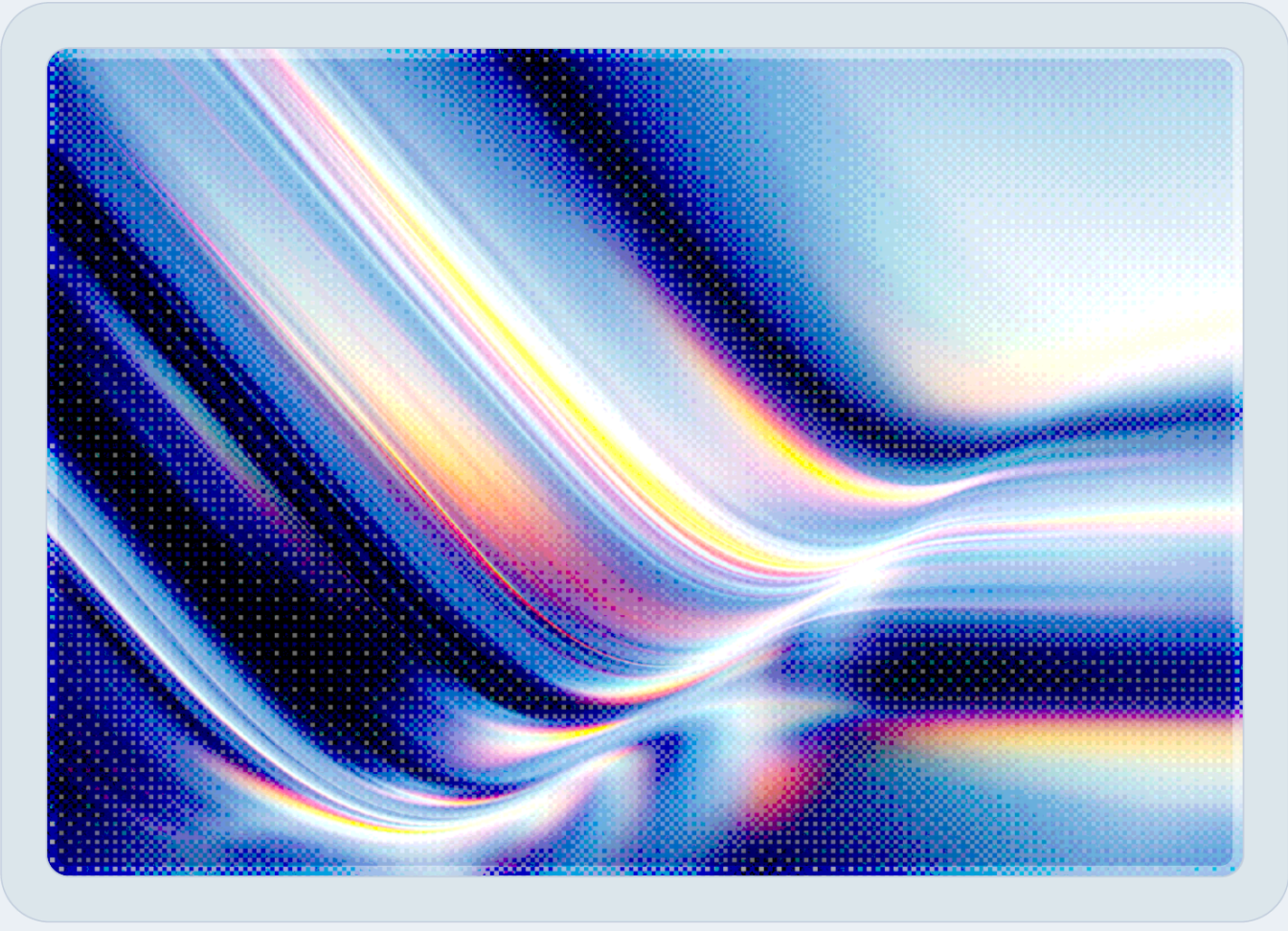
CSAI Foundation | Cloud Security Alliance

Project Glasswing: AI Discovery Outpaces Open Source Patching Capacity

1,596 Vulnerabilities Disclosed to Maintainers — Only 97 Fixed

2026-06-08

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- Project Glasswing, Anthropic's coordinated AI vulnerability research initiative, deployed Claude Mythos Preview alongside twelve major technology partners in April 2026 and identified over 10,000 high- or critical-severity vulnerabilities within its first month of operation. The program subsequently expanded to approximately 150 additional organizations in more than 15 countries [1][4][6].
- As of May 22, 2026, Anthropic had disclosed 1,596 of those findings to open-source maintainers across 281 projects; only 97 are confirmed patched — roughly a six percent remediation rate on disclosed findings and less than one percent of the total discovery volume [2][3].
- The primary security bottleneck has shifted from vulnerability *discovery* to vulnerability *remediation*: AI-powered research tools operate at machine speed while open-source maintainers, many of them volunteers, operate at human speed and are increasingly capacity constrained.
- Exploitation timelines have compressed sharply independent of this development: the median time from public disclosure to weaponized exploit has fallen from 771 days in 2018 to single-digit hours, and 28.3% of CVEs are now exploited within 24 hours of disclosure [5][3].
- The National Vulnerability Database is under structural strain, with CVE submissions rising 263% between 2020 and 2025, limiting enrichment to the highest-risk submissions and degrading the metadata that automated triage tools depend on [3].
- Organizations with dependencies on the 281 open-source projects in which Glasswing has disclosed vulnerabilities should treat those disclosures as emergency-priority items and audit their NVD-dependent triage tools for degraded signal quality due to the database's acknowledged processing backlog.

Background

Anthropic launched Project Glasswing on April 7, 2026, as a coordinated vulnerability research initiative built around Claude Mythos Preview, an unreleased frontier AI model designed for autonomous security research. Unlike prior AI-assisted security tools that functioned as analytical layers atop human review,

Mythos Preview is capable of autonomously identifying zero-day vulnerabilities, developing working exploits, and chaining multiple independent flaws into compound attack paths without human direction [1][4].

The model's performance on established benchmarks significantly exceeds current production AI models. Mythos Preview achieved 83.1% on CyberGym vulnerability-detection benchmarks, compared to 66.6% for Claude Opus 4.6, and demonstrated a 72.4% success rate on exploit development benchmarks — a capability that is effectively near-zero for current-generation production models [4][5]. Its autonomous capabilities include building multi-gadget return-oriented programming chains and chaining four independent vulnerabilities into a single attack path. Anthropic has withheld Mythos Preview from general access, citing dual-use risks, while extending access to a structured cohort of organizational partners under coordinated disclosure agreements [4][6].

Glasswing's launch partners included twelve major technology organizations: Amazon Web Services, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorganChase, the Linux Foundation, Microsoft, NVIDIA, and Palo Alto Networks, alongside Anthropic itself [4]. Anthropic committed $100 million in model usage credits to participating organizations, $2.5 million to Alpha-Omega and the Open Source Security Foundation (OpenSSF), and $1.5 million to the Apache Software Foundation to support open-source remediation infrastructure [4]. In a second wave announced in late May 2026, approximately 150 additional organizations in more than 15 countries joined the program, extending the initiative into power, water, healthcare, communications, and hardware sectors — environments where many high-severity vulnerabilities could affect more than 100 million people if exploited [6].

The name evokes the glasswing butterfly's transparent wings — a fitting metaphor for vulnerabilities that have persisted undetected for years or decades in widely deployed software, not because they were concealed, but because no sufficiently powerful analytical instrument had been applied to find them.

Security Analysis

The Scale and Validity of Glasswing Findings

The vulnerability discovery volume Project Glasswing has produced is on a scale with no clear public precedent in reported quantity and third-party-validated severity rate. Within the initiative's first month, partners reported more than 10,000 high- or critical-severity vulnerabilities across software they maintain or depend on [1]. Anthropic estimated an additional 6,202 high- or critical-severity

vulnerabilities in open-source code across more than 1,000 scanned projects beyond those directly reported [1][2]. Cloudflare alone identified approximately 2,000 vulnerabilities using Mythos Preview, of which 400 were rated critical or high [1][2].

These are not raw automated scanner outputs. Independent security firms assessed 1,752 of the findings and determined that 90.6% were valid, with 62.4% confirmed as high- or critical-severity [2]. This 90.6% validity rate represents a high signal-to-noise ratio that is significant for understanding the triage burden placed on maintainers: they are receiving primarily real, severe vulnerabilities, yet cannot process them at the rate of disclosure.

The age of some findings underscores how qualitatively different AI-powered discovery is from prior approaches. Mythos Preview found a 17-year-old remote code execution vulnerability in FreeBSD that allows an attacker to obtain root access on any machine running NFS — subsequently assigned CVE-2026-4747 [7]. It identified a 27-year-old vulnerability in OpenBSD enabling remote system crashes, and a 16-year-old flaw in FFmpeg that had evaded more than five million automated test executions [4]. These are foundational projects with extensive security review histories. The implication is not merely that these specific bugs existed, but that AI-powered discovery tools can now surface critical vulnerabilities in mature, heavily scrutinized software that decades of human review had missed — and can do so systematically and at scale.

The Structural Patching Deficit

Project Glasswing has exposed the discovery-to-remediation gap as a primary — and previously underappreciated — systemic security risk. As of May 22, 2026, Anthropic had disclosed 1,596 vulnerabilities to open-source maintainers across 281 projects. The ecosystem has formally acknowledged 88 of those findings through CVE or GHSA assignment — roughly 5.5 percent — and confirmed patches for 97, or roughly six percent, while thousands of additional validated vulnerabilities in proprietary codebases remain in the disclosure pipeline [2][3].

The causes of this deficit appear to be structural rather than motivational. By most published measures, open-source software is predominantly maintained by volunteers or small teams without dedicated security budgets, though projects under well-resourced foundations or major vendor stewardship represent important exceptions. Patching a complex vulnerability is not a single act: it requires understanding the root cause, designing a fix that avoids regressions, testing across configurations and platforms, coordinating a disclosure timeline, and communicating to downstream consumers. Several maintainers have told Anthropic they are severely capacity constrained and have asked the company to slow its disclosure cadence so they have more time to design patches [2]. That a well-resourced initiative

with coordinated partner access is straining the ecosystem's patching capacity suggests that less structured AI-powered discovery — including by adversaries — could produce a cascade for which the ecosystem has no absorption mechanism.

The challenge is further compounded by the deteriorating state of vulnerability metadata infrastructure. The National Vulnerability Database processed a 263% increase in CVE submissions between 2020 and 2025 and has acknowledged that it can now enrich only the highest-risk submissions [3]. Automated vulnerability management tools that rely on CVSS scores or NVD enrichment for triage will increasingly encounter unscored or incompletely scored entries, degrading signal quality precisely when the volume of actionable findings is at an all-time high.

Exploitation Timelines Have Compressed

The patching deficit is not a static risk tolerance problem — it is an accelerating one. The median time from public vulnerability disclosure to weaponized exploit has dropped from 771 days in 2018 to single-digit hours in recent years [5]. Threat intelligence analysis, citing Mandiant data, indicates that exploitation is now occurring, on average, before patches are available [8]. Independent analysis finds that 28.3% of CVEs are exploited within 24 hours of public disclosure [3]. The window in which a disclosed but unpatched vulnerability is exclusively a defender concern — rather than simultaneously an active attacker opportunity — has effectively closed.

This compression matters acutely in the context of Glasswing's disclosure volume. Coordinated disclosure, the standard practice of privately notifying vendors and allowing a remediation window before public announcement, rests on the assumption that the disclosure period is long enough for a patch to be developed, tested, and deployed before exploitation begins. If exploitation can occur within hours of public disclosure while patching takes weeks, the coordinated disclosure model provides diminishing assurance for many critical infrastructure operators, particularly those in power, water, and healthcare sectors where unplanned downtime carries direct safety consequences — organizations that cannot deploy emergency patches without regulatory approval or service disruption risks.

There is an additional systemic risk in the broader ecosystem of AI-generated vulnerability reports. The curl project shut down its public bug bounty program after AI-generated submissions overwhelmed reviewers with unverified findings, and FFmpeg maintainers have publicly characterized AI-generated reports as "CVE slop" [9]. High-quality disclosures from initiatives like Glasswing may be deprioritized in maintainer queues because they arrive alongside a flood of low-quality AI-generated noise — a tragedy-of-the-commons dynamic that could, in aggregate, make AI's net contribution to open-source security negative if AI-generated noise degrades triage capacity faster than high-quality AI-assisted discovery reduces risk.

Asymmetric Risk for Critical Infrastructure

Glasswing's second cohort expansion into power, water, healthcare, and communications sectors reflects recognition that the patching deficit is most acute in environments with the longest remediation cycles [6]. Industrial control systems, medical devices, and communications infrastructure operate on update cadences measured in quarters or years, driven by regulatory certification requirements, service continuity constraints, and the high cost of downtime. A critical vulnerability in a hospital DICOM server or power grid management system that cannot be patched for many months or longer represents a categorically different risk profile than an unpatched web application library: the consequence window is longer, the compensating control options are more limited, and the aggregate impact of exploitation is larger.

In one documented AI-assisted campaign from early 2026, a threat actor used AI-orchestrated tools to scan 2,516 FortiGate targets across 106 countries, compromising more than 600 devices [2]. For critical infrastructure operators, the appropriate framing is not solely "how quickly can we patch?" but equally "what compensating controls can we deploy immediately to reduce exposure during the patching window?" — a question that requires pre-planned response procedures rather than reactive improvisation.

Recommendations

Immediate Actions

Organizations that maintain or depend on open-source software should immediately audit their direct and transitive dependencies against the 281 open-source projects in which Glasswing has disclosed vulnerabilities, prioritizing those that are internet-facing or that handle authentication, cryptography, or network protocol parsing. Where patches are available — the 97 confirmed remediations represent verified, high-confidence fixes — they should be treated as emergency changes rather than scheduled maintenance items, given the documented exploitation timelines in the current threat environment.

Security teams should audit their vulnerability management workflows for NVD dependency. Tools that rely on CVSS scores or NVD enrichment as the primary triage signal are increasingly unreliable given the database's acknowledged processing backlog. Supplementing or replacing NVD-based scoring with GitHub Security Advisories (GHSA), direct vendor security advisories, and established private advisory channels provides a more current and reliable signal.

Critical infrastructure operators who cannot immediately patch vulnerable components should implement network segmentation to isolate affected systems, apply least-privilege access controls to reduce lateral movement opportunity, and deploy comprehensive logging to detect exploitation attempts. The objective is to reduce the blast radius of potential exploitation during the mandatory patching window.

Short-Term Mitigations

Over the next three to six months, security teams should establish direct communication channels with the maintainers of their most critical open-source dependencies — using GitHub's private security advisory feature or comparable mechanisms — so that coordinated disclosures are received before public announcement. This effectively extends the response window for organizations that would otherwise learn of a vulnerability when the CVE is published, which may be hours before exploitation is documented in the wild.

Security operations teams should review their detection rules for the vulnerability classes Glasswing has most frequently documented: memory corruption, privilege escalation via kernel interfaces, remote code execution through network-facing services, and logic flaws in authentication and cryptographic libraries. AI-powered vulnerability discovery tends to identify entire families of related flaws rather than isolated instances, meaning a disclosed CVE in one component is a signal to audit the same pattern across related codebases.

Organizations should begin evaluating AI-assisted patch validation tooling to compress the validation cycle between patch development and deployment. The bottleneck is not exclusively patch development but also the validation that a fix does not introduce regressions across configurations — a step that AI-powered testing can materially accelerate. Anthropic's Claude Security product represents one available option for initiating this evaluation, among other available AI-assisted code scanning tools including GitHub Copilot Autofix and Semgrep Assistant [6].

Strategic Considerations

Project Glasswing's deeper implication is that the security industry's established operating model — discover vulnerabilities, disclose them, wait for vendors to patch, then deploy — was designed for conditions in which discovery was slow and attacker capability was human-bounded. Neither condition holds. Organizations that treat vulnerability management as a compliance function, executing scheduled scan cycles on fixed cadences, are running a process calibrated for a threat environment that no longer exists. Security leaders should treat Glasswing as a leading indicator: any autonomous vulnerability

discovery capability available to Anthropic's partners today will eventually become accessible to adversaries, and the patching capacity deficit must be addressed structurally before that symmetry arrives.

The long-term response requires sustained investment in the open-source maintainer ecosystem as shared critical infrastructure. Anthropic's $2.5 million commitment to Alpha-Omega and OpenSSF establishes a precedent, but the structural problem is that the world's most widely deployed software depends on largely uncompensated labor. Large technology consumers of open-source software, industry consortia, and government agencies should collectively treat maintainer capacity as an infrastructure problem with a corresponding infrastructure investment model — analogous to the commitment to physical infrastructure security that followed major incidents demonstrating the cost of underinvestment.

Anthropic's proposal for a third-party, independent body to coordinate large-scale AI-powered cybersecurity initiatives across public and private sectors deserves serious consideration [4]. No single organization has sufficient visibility or authority to govern AI-powered vulnerability disclosure at the scale Glasswing has demonstrated is now operationally possible. A coordinated governance body — with defined roles for AI developers, vulnerability discoverers, open-source maintainers, downstream software consumers, and government agencies — would provide the institutional infrastructure that current ad hoc coordination mechanisms cannot.

CSA Resource Alignment

Project Glasswing's findings engage directly with several CSA frameworks that provide actionable governance and technical control guidance for organizations responding to this threat landscape.

The **MAESTRO** framework for agentic AI threat modeling provides the foundational analysis for understanding how autonomous vulnerability research systems like Mythos Preview differ from prior AI-assisted tools in their governance requirements. MAESTRO's Layer 4 (memory and context management) and Layer 6 (integration layer) analysis apply directly to AI systems that conduct autonomous security research at scale, particularly regarding auditability of model decision-making, constraints on autonomous action, and human oversight mechanisms for consequential outputs. Organizations deploying AI-powered security tooling should use MAESTRO to model the specific threat surfaces introduced by agentic operation before deployment.

The **AI Controls Matrix (AICM)**, as a superset of the Cloud Controls Matrix (CCM), addresses governance requirements for AI systems operating in sensitive domains, including accountability, transparency, and human oversight controls applicable to autonomous vulnerability research. AICM's

supply chain security domain is directly relevant to the open-source dependency risk Glasswing has exposed: organizations should apply AICM controls to govern not only their own AI deployments but the AI-assisted security tooling embedded throughout their software supply chains, including the provenance and validation standards applied to AI-generated vulnerability reports.

The **STAR for AI** program and the AI Consensus Assessment Initiative Questionnaire (AI-CAIQ) provide vendor assessment mechanisms directly applicable to evaluating AI-powered security tools. As AI-generated vulnerability reports become widespread, procurement teams should incorporate AI-CAIQ criteria into the evaluation of any security vendor whose findings are produced or triaged by AI systems, including assessment of training data provenance, validation methodology, and responsible disclosure practices.

CSA's **Zero Trust** guidance remains the foundational compensating control framework for the window between vulnerability disclosure and patch deployment. The assumption that internal network traffic is trustworthy becomes untenable when exploitation timelines are measured in hours and patch cycles run for weeks or months. Zero Trust principles — verify explicitly, use least privilege access, assume breach — provide the architectural posture necessary for organizations operating in the current threat environment regardless of their patch deployment velocity.

References

[1] Anthropic. "Project Glasswing: An Initial Update." Anthropic, May 2026.

[2] The Hacker News. "Project Glasswing Proved AI Can Find the Bugs. Who's Going to Fix Them?." The Hacker News, April 2026.

[3] Cloud Security Alliance Labs. "Project Glasswing and the AI Vulnerability Disclosure Velocity Crisis." CSA Labs, 2026.

[4] Anthropic. "Project Glasswing: Securing Critical Software for the AI Era." Anthropic, April 2026.

[5] Penligent. "Project Glasswing and Claude Mythos Show the New AI Security Bottleneck." Penligent, 2026.

[6] Anthropic. "Expanding Project Glasswing." Anthropic, May 2026.

[7] Penligent. "CVE-2026-4747: FreeBSD RPCSEC_GSS Remote Code Execution." Penligent, 2026.

[8] Chainguard. "AI Is Finding Vulnerabilities Faster Than Anyone Can Patch Them. Now What?." Chainguard, 2026.

[9] Security Boulevard. "AI Vulnerability Discovery and the Open Source CVE Surge." Security Boulevard, May 2026.