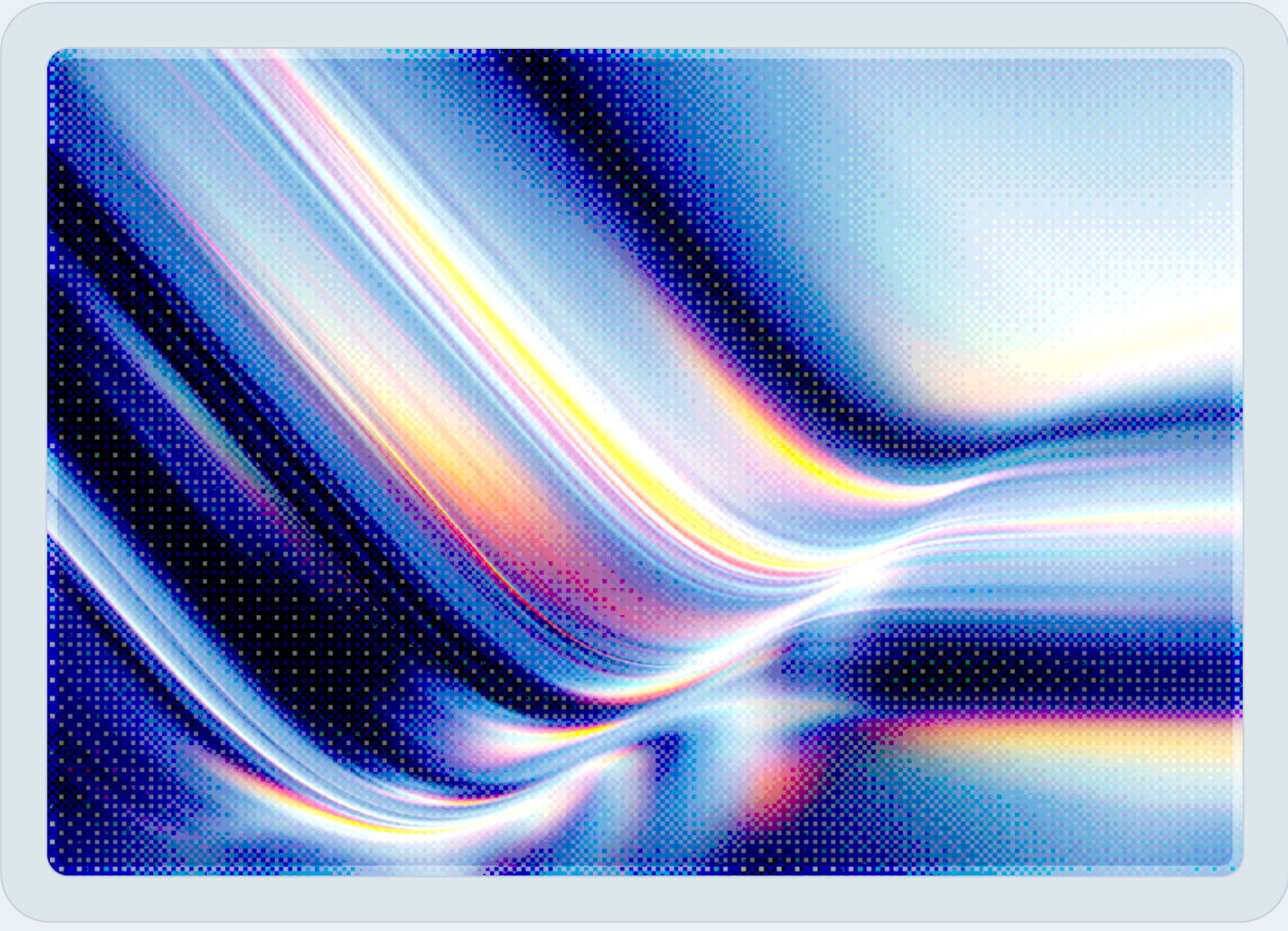
CSAI Foundation | Cloud Security Alliance

Project Glasswing and the AI Vulnerability Disclosure Velocity Crisis

How Autonomous Vulnerability Discovery Has Outpaced the Global Capacity to Patch

2026-05-24

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- Anthropic's Project Glasswing — a coalition of twelve major technology organizations using Claude Mythos Preview for autonomous vulnerability research — discovered more than ten thousand high- and critical-severity vulnerabilities across systemically important open-source software in its first month of operation, demonstrating that AI can now conduct large-scale, autonomous zero-day discovery at operational scale.
- Of approximately 1,596 vulnerabilities Anthropic has disclosed to open-source maintainers as of May 2026, only 97 have been patched — roughly a six percent remediation rate — exposing a structural mismatch between the velocity at which AI can surface flaws and the pace at which the global developer community can remediate them.
- NIST acknowledged that the National Vulnerability Database can no longer keep pace with incoming CVE volume, announcing in April 2026 that it will limit enrichment to only the highest-risk submissions; CVE submissions increased 263 percent between 2020 and 2025, and the pace is accelerating.
- Time-to-exploit has inverted: Mandiant's M-Trends 2026 report found that 28.3 percent of CVEs are now exploited within 24 hours of public disclosure, and threat actors have demonstrated the ability to weaponize newly published vulnerabilities faster than most organizations can deploy patches.
- The traditional ninety-day coordinated disclosure window was designed for human-paced research; AI-paced discovery at tens of thousands of findings per month requires new disclosure governance models that account for maintainer capacity, aggregate risk, and prioritized remediation sequencing.
- Security teams should immediately review their vulnerability intake and triage processes, prioritize AI-enabled patch validation tooling, and engage with FIRST's Vulnerability Exploitability eXchange (VEX) standard, OpenSSF's disclosure guidance, and other emerging disclosure frameworks rather than relying solely on the CVE/NVD pipeline, which is now structurally overwhelmed.

Background

On April 7, 2026, Anthropic launched Project Glasswing, a coordinated vulnerability research initiative built around Claude Mythos Preview — an unreleased AI model designed for autonomous security research. The coalition brought together Amazon Web Services, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorgan Chase, the Linux Foundation, Microsoft, NVIDIA, and Palo Alto Networks, along with approximately fifty partner organizations collectively granted controlled access to Mythos for defensive vulnerability discovery. [1, 2]

Claude Mythos Preview achieved an 83.1 percent success rate on the CyberGym cybersecurity vulnerability reproduction benchmark — a performance level that Anthropic characterized as exceeding all but the most skilled human security researchers at finding and exploiting software vulnerabilities. [2] That benchmark figure is not merely a competitive metric: it indicates that the model can reliably reproduce conditions necessary to confirm that a reported flaw is genuine, which is the most labor-intensive step in vulnerability triage. The practical effect is that Mythos can apply expert-level analysis across the breadth of systemically important software simultaneously, rather than being constrained to the subjects a single human or team can cover in any given period; the Glasswing operational results provide the confirmation that this benchmark-level performance translates to real-world discovery at scale.

Within one month of launch, Project Glasswing participants collectively reported more than ten thousand high- or critical-severity vulnerabilities, with Anthropic estimating an underlying discovery count of approximately 23,019 total vulnerabilities including 6,202 rated high or critical severity. [3, 6] Of those, 1,726 have been validated as confirmed true positives and 1,094 confirmed as high- or critical-severity flaws. [3] Among the specific discoveries were a 27-year-old vulnerability in OpenBSD and a 16-year-old bug in FFmpeg — both of which had gone undetected across decades of human review and automated static analysis. [1, 16] Cloudflare, one of the coalition partners, reported discovering 2,000 bugs in its own critical-path systems, 400 of them high or critical severity, and characterized Mythos' false positive rate as better than that of human testers. [1]

The significance of those findings extends well beyond their individual severity ratings. The OpenBSD and FFmpeg cases illustrate that AI vulnerability research is not merely replicating what human researchers already do faster — it is surfacing entire categories of latent exposure that decades of conventional security practice left undiscovered. Software that has been in continuous production use and regular security review for nearly three decades can harbor critical flaws that only become visible when a sufficiently capable model examines the code with sustained attention at a scale no human team could maintain.

Security Analysis

The Patch Asymmetry

The central security challenge Project Glasswing has illustrated at operational scale is not the discovery of vulnerabilities — it is the capacity to address them. As of May 22, 2026, Anthropic had directly disclosed 1,596 vetted findings to maintainers of 281 open-source projects, of which 97 had been patched and 88 had received a CVE or GitHub Security Advisory assignment. [6] As of that date, only six weeks had elapsed since the first disclosures began flowing through 90-day windows, and some portion of the low patch rate reflects timing as much as capacity — many windows are simply not yet due. Even accounting for that timing effect, the pace already suggests a structural gap that open-source developers themselves have characterized as unsustainable. [4, 13]

Open-source software is overwhelmingly maintained by individuals and small teams working under significant resource constraints. Vulnerability remediation is not a simple matter of accepting a disclosure and writing a fix. A maintainer who receives a critical vulnerability report must first reproduce and validate the finding, understand its scope and attack surface, design a patch that resolves the root cause without introducing regressions, test that patch across supported configurations, coordinate a release, and draft a public advisory — all before the disclosure window expires. For a project with a single active maintainer, this process can realistically take weeks even for a straightforward flaw. When that maintainer faces a queue of dozens of such reports, each with its own 90-day clock, the arithmetic becomes untenable. Some open-source developers have reportedly asked Anthropic directly to slow its disclosure rate because the current pace is simply not sustainable against their available bandwidth. [4, 13]

The disclosure architecture Anthropic published for Project Glasswing attempts to account for this reality. The default timeline is ninety days after notification or upon patch release, matching the long-standing Google Project Zero standard. For actively exploited critical vulnerabilities the window compresses to seven days. [5] Anthropic also maintains a public disclosure dashboard tracking the status of each reported vulnerability, providing transparency into the remediation pipeline. [6] These design choices are reasonable within a conventional coordinated disclosure framework, but they do not resolve the underlying capacity problem: when findings arrive faster than the recipient community can process them, even well-designed disclosure windows do not prevent a growing backlog of unpatched, known-but-not-yet-public vulnerabilities.

The NVD Breaks Under Load

The National Vulnerability Database has been the backbone of enterprise vulnerability management for two decades: security tools depend on NVD enrichment — CVSS scores, CWE classifications, CPE mappings — to drive automated triage, patch prioritization, and compliance reporting. That infrastructure is now showing signs of structural failure under the weight of AI-accelerated discovery. CVE submissions increased 263 percent between 2020 and 2025, and NIST acknowledged in April 2026 that it can no longer enrich the full volume of incoming vulnerabilities. [7] Going forward, NIST will apply detailed analysis only to the highest-risk submissions; every unenriched CVE published before March 1, 2026 has been reclassified as "Not Scheduled," acknowledging that the existing backlog cannot be cleared at the current submission rate. [7]

The operational consequence is that the vast majority of new vulnerabilities will now enter the CVE ecosystem without the CVSS metadata that automated tooling requires for downstream prioritization. Vulnerability management platforms that depend on NVD enrichment will be making triage decisions with incomplete or absent severity data. For organizations that have not built supplemental enrichment pipelines, this means a growing proportion of their vulnerability inventory will be effectively invisible to automated risk scoring — precisely when the volume of relevant findings is at a historic high.

Time-to-Exploit Has Inverted

The threat actors operating on the other side of this equation are not constrained by the same capacity limitations as defenders. Mandiant's M-Trends 2026 report found that 28.3 percent of disclosed CVEs are now being exploited within 24 hours of public disclosure, and time-to-exploit across the vulnerability population has compressed from 63 days in 2018 to approximately 5 days in recent measurement periods. [8] CrowdStrike's 2026 Global Threat Report observed a 42 percent year-over-year increase in zero-days exploited before any public disclosure in 2025, indicating that threat actors are not always waiting for CVE publication to begin weaponization. [9]

The practical implication is that the ninety-day disclosure window — a policy designed to give defenders an advantage over attackers — is being eroded from both ends simultaneously. On one end, AI-accelerated discovery generates more findings than the remediation pipeline can absorb, creating a growing backlog of known vulnerabilities without available patches. On the other end, threat actors with access to similar AI capabilities can identify and exploit vulnerabilities before patches are available regardless of disclosure timeline. The gap between disclosure and exploitation is closing to hours in the most urgent cases, while the gap between discovery and patch availability widens.

The same autonomous discovery capability that Project Glasswing deploys for defensive purposes is increasingly accessible to adversarial actors operating without disclosure obligations. The Mandiant and CrowdStrike data cited above — 28.3 percent of CVEs exploited within 24 hours and a 42 percent year-over-year acceleration in pre-disclosure zero-day weaponization — already reflect, at least in part, the effect of AI-assisted offensive research. Glasswing-class tools in adversarial hands do not produce a disclosure backlog; they produce a weaponization pipeline with no public notification step. Governance frameworks designed to manage AI-paced disclosure must therefore account for both vectors: the structural question of how defenders communicate and remediate at scale, and the parallel reality that adversaries may have already completed their discovery and weaponization cycle before any CVE is published.

The Triage Fatigue Cascade

The volume problem extends beyond open-source maintainers to the broader vulnerability disclosure ecosystem. HackerOne suspended new submissions to its Internet Bug Bounty program effective March 27, 2026, citing an inability to maintain review capacity against the volume of AI-generated reports; the program's valid submission rate had fallen from approximately fifteen percent to below five percent as AI-generated findings flooded the intake queue. [10] This pattern — where AI tooling generates submissions at a volume that overwhelms human review capacity, including for well-resourced vulnerability disclosure platforms — is a preview of the organizational challenge that security teams managing their own vulnerability intake programs will increasingly face.

DARPA's AI Cyber Challenge, which concluded at DEF CON 33 in August 2025, demonstrated that fully autonomous AI systems could collectively analyze more than 54 million lines of code across 53 challenge software projects, reproduce 63 verified challenge problem vulnerabilities, and discover 25 previously unknown real-world flaws at an average cost of roughly $152 per task. [11, 14] The cost and time parameters from that competition are approaching the range where adversarial actors with modest resources could deploy comparable capability against targets of interest. The question for enterprise security teams is no longer whether AI-scale vulnerability discovery will affect their environment — Project Glasswing has confirmed that it already is — but whether their triage, patch, and detection architectures can respond at a comparable rate.

Recommendations

Immediate Actions

Security teams should audit their current vulnerability ingestion pipeline for NVD dependency and identify which triage and prioritization workflows will degrade as NVD enrichment coverage narrows. Where CVSS scores from NVD are used to gate automated patch deployment or compliance reporting, those workflows need a supplemental enrichment source — commercial threat intelligence feeds, EPSS (Exploit Prediction Scoring System) data, or vendor-specific advisories — to maintain triage fidelity. Teams should also verify that their vulnerability management platforms are consuming GitHub Security Advisory data (GHSA) in addition to CVE records, as an increasing share of Project Glasswing disclosures will enter the ecosystem through GHSA before or instead of CVE assignment.

Organizations running open-source software at scale should identify their most critical-path dependencies and establish direct notification relationships with maintainer projects through platforms like GitHub's private security advisory feature. Waiting for NVD enrichment to surface vulnerabilities in critical open-source components is no longer a viable detection strategy; the combination of slower enrichment and faster exploitation means the signal will arrive too late for many organizations relying solely on that channel.

Short-Term Mitigations

Security engineering teams should evaluate AI-assisted patch validation tooling as a complement to existing patch management workflows. OpenAI's Daybreak program, announced in May 2026, represents one approach to automating patch verification — confirming that a proposed fix actually resolves the reported flaw without introducing new weaknesses. [12] Broader adoption of AI-assisted patch validation is likely necessary to close the gap between discovery velocity and remediation capacity; tools that can reduce the human effort required to validate a candidate patch will be critical to improving the patch rate against AI-generated findings. [15]

Organizations that consume open-source software should also evaluate their contribution posture toward the projects they depend on most. The maintainer capacity crisis that Project Glasswing has surfaced is not solely a problem for the organizations receiving disclosures; every consumer of open-source software has an organizational interest in the viability of the maintainer ecosystem. Funding through Open Source Security Foundation (OpenSSF) programs, contributing engineering time for security-focused patch review, and participating in OpenSSF's Secure Supply Chain Consumption Framework are concrete ways organizations can contribute to closing the patch velocity gap.

Strategic Considerations

The coordinated vulnerability disclosure framework was designed in an era when discovery velocity was bounded by human researcher capacity. The events of April and May 2026 provide the strongest empirical case yet that this assumption no longer holds at AI-paced discovery rates. Industry stakeholders — including AI developers, platform vendors, open-source foundations, and standards bodies — will need to develop disclosure governance models that can operate at AI-paced discovery rates. Tiered disclosure timelines based on maintainer capacity and aggregate risk, rather than fixed per-vulnerability windows, may be necessary. Centralized patch triaging infrastructure — analogous to the role CISA's Known Exploited Vulnerabilities catalog plays for prioritization — could help focus remediation resources where exploitation risk is highest.

For organizations with internet-facing systems running affected open-source components, runtime detection, behavioral anomaly monitoring, and network-layer controls that can identify exploitation attempts independent of patch status are essential backstops in an environment where patch lag is structurally unavoidable for a meaningful proportion of the vulnerability inventory. Given that time-to-exploit for critical vulnerabilities can now be measured in hours, the window in which organizations can rely on a vulnerability being unknown to adversaries after its disclosure is effectively closed. These detection-oriented compensating controls are not optional supplements; they are load-bearing elements of any security posture that depends on open-source software.

CSA Resource Alignment

Project Glasswing and the surrounding velocity crisis intersect directly with several areas of active Cloud Security Alliance guidance. The AICM (AI Controls Matrix) addresses the security governance implications of AI systems operating with autonomous capability — the model trust, containment, and disclosure governance questions raised by Mythos-class systems are precisely the domain the AICM's AI supply chain and AI operational security controls address. Organizations evaluating whether to participate in AI-powered vulnerability research coalitions can use the AICM's shared responsibility model to assign accountability for vulnerability intake, validation, and disclosure decisions across their application provider, orchestration, and infrastructure layers.

MAESTRO, CSA's threat modeling framework for agentic AI systems, is directly applicable to the autonomous vulnerability discovery use case. Claude Mythos Preview is an agentic system operating with significant autonomy over codebase analysis and exploit reproduction — exactly the class of AI agent MAESTRO was designed to evaluate for containment failures, trust boundary violations, and

unintended side effects. Organizations building internal AI-assisted vulnerability research programs should use MAESTRO to threat-model their AI agent's access to production code, findings databases, and disclosure workflows before deployment.

The STAR for AI program and the AI-CAIQ provide an assessment and assurance framework relevant to the disclosure architecture question: organizations receiving AI-generated vulnerability reports from coalition partners or commercial AI security tools should be able to assess the provenance, validation methodology, and false-positive rate of those findings as part of their triage process. STAR for AI attestations could eventually serve as a trust signal in the disclosure pipeline, providing recipients with documented assurance about the discovery AI's operational characteristics before they commit remediation resources.

CSA's Zero Trust guidance applies to the compensating control posture recommended above. In an environment where patch availability consistently lags behind exploitation, network segmentation, micro-perimeter enforcement, and workload-level access controls that can limit the blast radius of exploitation – independent of whether a patch is available – become primary security controls rather than supplementary defense-in-depth measures for any organization running exposed open-source components at scale.

References

[1] Anthropic. "Project Glasswing: An initial update." Anthropic, May 2026.

[2] Anthropic. "Claude Mythos Preview." Anthropic Red Team, April 2026.

[3] The Next Web. "Anthropic's Claude Mythos found 10,000 critical vulnerabilities in one month. The patches can't keep up.." The Next Web, May 2026.

[4] Security Affairs. "Anthropic's Glasswing: 10,000+ Vulnerabilities Found in One Month, and the Patching Problem Has Never Been More Obvious." Security Affairs, May 2026.

[5] Bloo. "Project Glasswing: How Anthropic's Disclosure Architecture Actually Works." Bloo, 2026.

[6] Let's Data Science. "Anthropic Publishes Coordinated Vulnerability Disclosure Dashboard." Let's Data Science, 2026.

[7] NIST. "NIST Updates NVD Operations to Address Record CVE Growth." NIST, April 2026.

[8] Google Cloud / Mandiant. "M-Trends 2026." Mandiant, 2026.

[9] CrowdStrike. "2026 CrowdStrike Global Threat Report." CrowdStrike, 2026.

[10] Dark Reading. "AI-Led Remediation Crisis Prompts HackerOne to Pause Bug Bounties." Dark Reading, 2026.

[11] DARPA / arXiv. "SoK: DARPA's AI Cyber Challenge (AIxCC): Competition Design, Architectures, and Lessons Learned." arXiv, 2026.

[12] The Hacker News. "OpenAI Launches Daybreak for AI-Powered Vulnerability Detection and Patch Validation." The Hacker News, May 2026.

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

[14] Resilient Cyber. "Vulnpocalypse: AI, Open Source, and the Race to Remediate." Resilient Cyber, 2026.

[15] Chainguard. "AI is finding vulnerabilities faster than anyone can patch them. Now what?." Chainguard, 2026.

[16] VentureBeat. "Mythos autonomously exploited vulnerabilities that survived 27 years of human review. Security teams need a new detection playbook." VentureBeat, 2026.