CSAI Foundation

Cloud Security Alliance AI Safety Initiative

The Collapsing Exploit Window

Systemic Consequences of AI-Autonomous Vulnerability Discovery at Scale

Unofficial AI-assisted Research

2026-04-23

Table of Contents

Executive Summary

For decades, the security industry operated on an implicit assumption: the window between a vulnerability's discovery and its weaponization was wide enough, on average, to permit organized response. Vendors would patch, defenders would deploy, and the cycle would continue at a pace that human institutions — disclosure programs, patch management processes, vulnerability databases — could track and support. That assumption no longer holds.

In April 2026, Anthropic disclosed that its Claude Mythos Preview model had autonomously identified thousands of previously unknown, high-severity vulnerabilities across every major operating system, every major web browser, and a range of critical libraries — including flaws that had persisted, undetected, for 17 to 27 years [1][20]. The 271 zero-day vulnerabilities discovered in Mozilla Firefox alone [2] represent a volume that would have taken a large human research team years to produce. Claude Mythos accomplished it at a cost of under $50 per codebase survey [3]. Anthropic's disclosure represented not an isolated research achievement but a capability crossing — the moment when AI systems surpassed all but the most elite human researchers at systematic vulnerability discovery and exploitation.

This capability crossing arrives at the worst possible moment. The proportion of exploited vulnerabilities weaponized within 24 hours of disclosure has climbed to 28.3% as of 2025, and 67.2% of exploited CVEs are now abused on or before their public disclosure date — meaning defenders receive no advance warning at all [4]. The window that once provided organized response time has in many cases ceased to exist. The global infrastructure designed to manage vulnerability information — the National Vulnerability Database, coordinated disclosure programs, patch distribution channels — is already failing under the weight of record-breaking CVE volume, with CVE submissions having increased 263% between 2020 and 2025 [5]. NIST formally acknowledged in April 2026 that it can no longer enrich all CVEs, retreating to a triage model that leaves a large fraction of disclosed vulnerabilities without structured analysis [6].

The convergence of these trends is not merely a technical problem. It represents a structural shift in the economics of cyber offense and defense — one that erodes the institutional foundations of coordinated security response and concentrates operational advantage with those who move fastest. This paper examines the mechanisms of that shift, the cascading failures it is producing across the vulnerability management ecosystem, and the strategic adjustments that organizations and the broader security community must make to remain viable under conditions that the current security architecture was never designed to handle.

Introduction: The Architecture of a Vanishing Window

The concept of the exploit window refers to the period between a vulnerability's existence becoming known — whether through public disclosure, researcher discovery, or vendor notification — and the point at which defensive measures are broadly deployed. That window has always been contested territory. Attackers raced to develop exploits; defenders raced to develop and distribute patches. The coordination machinery of responsible disclosure — CVE identifiers, CVSS scoring, NVD enrichment, CISA advisories — emerged specifically to give that race intelligible structure, ensuring that defenders could prioritize intelligently and that vendors had fair warning before attacks began.

The machinery was never perfect. High-profile vulnerabilities like Log4Shell demonstrated that even well-resourced organizations struggled to inventory and patch within reasonable timeframes. Exploitation of zero-days by nation-state actors showed that some vulnerabilities never received a defensive head start at all. But the system functioned well enough, for long enough, that the security industry built its operating assumptions around it. Patch cadences were measured in weeks. Mean time to remediate was measured in days for critical findings. The machinery, however imperfect, had a recognizable shape.

What has changed is not merely the speed of exploitation — though that is dramatic and well-documented — but the economics and scale of vulnerability discovery itself. AI systems have made the discovery of exploitable vulnerabilities radically cheaper, faster, and more accessible, while simultaneously producing a volume of findings that the existing coordination infrastructure cannot absorb. The result is a system under pressure at both ends: discovery accelerating while remediation capacity stagnates or erodes.

This paper focuses on three interlocking dimensions of that problem. The first is empirical: what AI vulnerability research systems can now demonstrably accomplish, and what the current data on exploitation timelines reveals about the structural consequences. The second is institutional: how the organizations and processes that coordinate vulnerability response are failing under the resulting load. The third is strategic: what the convergence of these trends implies for enterprise security programs, policy, and the long-term architecture of the security ecosystem.

A New Generation of AI Vulnerability Research Systems

The current generation of AI vulnerability research tools represents a qualitative departure from earlier automated security testing approaches. Fuzzing, static analysis, and symbolic execution have been mainstays of vulnerability research for decades, and AI-augmented versions of those techniques produced meaningful results in controlled settings. What has changed with frontier large language models is the ability to reason about code semantics, chain vulnerabilities across abstraction boundaries, and operate without domain-specific scaffolding across diverse codebases — capabilities that begin to approximate the generalist reasoning of skilled human researchers.

Google's Project Zero published the first significant evidence of this shift in October 2024, when its Naptime research framework — later evolved into the Big Sleep project — discovered a real-world memory corruption vulnerability in SQLite [7]. The finding was significant not because of the vulnerability itself but because of how it was found: an LLM agent operated autonomously through a cycle of hypothesis generation, code inspection, and validation, identifying a flaw that had escaped prior automated analysis. By mid-2025, Big Sleep had found 13 distinct vulnerabilities in FFmpeg [8], and Google's AI-augmented OSS-Fuzz system had identified 26 vulnerabilities across open-source projects, including a flaw in OpenSSL that had remained undetected for two decades [9].

These results, impressive in isolation, were dwarfed by what Trend Micro's ÆSIR platform demonstrated beginning in late 2025. ÆSIR — which Trend Micro describes as combining MIMIR for real-time threat intelligence correlation with FENRIR for autonomous zero-day discovery — identified 21 critical CVEs across industry platforms including NVIDIA, Tencent, MLflow, and MCP tooling within its first months of operation [10]. The platform operates with human researchers directing the overall research program and validating findings, but the discovery work itself is described as largely autonomous [10].

The April 2026 Anthropic disclosure made clear how far these capabilities have advanced in a short period. Claude Mythos Preview represents a general-purpose frontier model applied to vulnerability research under a controlled program that Anthropic calls Project Glasswing. The breadth and depth of the findings distinguished Mythos from prior research programs in both scope and per-finding cost. The model identified thousands of high and critical-severity vulnerabilities across the full spectrum of commonly used operating systems and web browsers [1]. Within Firefox alone, it discovered 271 previously unknown vulnerabilities [2]. The age distribution of its findings was particularly revealing: a 27-year-old flaw in OpenBSD — an operating system maintained specifically for its security properties — a 17-year-old remote code execution vulnerability in FreeBSD's NFS implementation (assigned CVE-2026-4747), and a 16-year-old flaw in FFmpeg's H.264 codec all fell to Mythos's analysis [1][21].

The model's exploitation capabilities proved as significant as its discovery capabilities. Where Claude Opus 4.6 achieved only two working exploits in hundreds of attempts against Firefox vulnerabilities, Mythos Preview succeeded 181 times [3]. It autonomously developed sophisticated exploitation techniques including JIT heap sprays, return-oriented programming chains, and multi-vulnerability sandbox escapes — the class of techniques previously concentrated in elite nation-state offensive teams and sophisticated criminal groups. Critically, it achieved these results at a cost that fundamentally changes the economics of vulnerability research: under $2,000 per kernel exploit, under $50 per codebase survey [3].

The UK AI Security Institute independently validated Mythos's capabilities, confirming that the model completed end-to-end simulated 32-step network attacks and solved 73% of expert-level CTF challenges [3]. Anthropic's own assessment noted that the model had crossed what the security research community sometimes calls the "autonomous offensive threshold" — the point at which an AI system can conduct the full vulnerability research lifecycle, from initial code analysis through working exploit delivery, without meaningful human involvement in individual steps.

Earlier academic research anticipated this trajectory. A 2024 paper published on arXiv demonstrated that GPT-4 agents could autonomously exploit 87% of one-day vulnerabilities in real-world CVE test environments when provided with CVE descriptions, compared to 0% for all other tested models [11]. A follow-up study showed that teams of LLM agents could exploit zero-day vulnerabilities, with hierarchical planning agents directing specialized subagents to achieve results 4.3 times better than prior single-agent frameworks [12]. These results, achieved with models that have since been superseded by multiple generations of development, established the research baseline that Mythos has now substantially exceeded.

System	Organization	Key Findings	Discovery Mode
Big Sleep	Google Project Zero	13 FFmpeg CVEs, SQLite zero-day (CVE-2025-6965)	Semi-autonomous, human-directed
OSS-Fuzz AI	Google	26 OSS vulnerabilities including 20-year-old OpenSSL flaw	AI-augmented fuzzing
ÆSIR (MIMIR/FENRIR)	Trend Micro	21 critical CVEs across NVIDIA, Tencent, MLflow, MCP	Autonomous with human validation
Mythos Preview	Anthropic	Thousands of zero-days across all major OS/browsers; 271 in	Fully autonomous agentic

System	Organization	Key Findings	Discovery Mode
		Firefox	

What Table 1 illustrates is not a single breakthrough but a trajectory. Each successive system demonstrates greater autonomy, broader scope, and lower per-finding cost than its predecessors. The practical implication is that vulnerability discovery is transitioning from a labor-intensive craft requiring rare human expertise to a computational task that can be parallelized across codebases, automated at scale, and repeated continuously. That transition has profound consequences for the entire vulnerability management ecosystem.

The Mathematics of the Collapsed Window

To appreciate the severity of what AI-accelerated discovery implies for defenders, it is necessary to understand how dramatically exploitation timelines have already compressed — before AI vulnerability discovery became a mass-market capability.

The historical baseline is stark. In 2018, the mean time between a vulnerability's public disclosure and its confirmed exploitation in the wild was approximately 756 days [4]. This figure, while imperfect as a measure of defender safety (since many exploits operated before public disclosure), at least indicated that defenders had meaningful time to respond in the majority of cases. By 2025, the picture had transformed: 28.3% of exploited vulnerabilities were weaponized within 24 hours of disclosure, and 56% were weaponized within the first month [4][24]. Most consequentially, the proportion of exploited CVEs that were abused on or before their public disclosure date — meaning defenders received no warning at all — climbed from 16.1% in 2018 to 67.2% by 2026 [4].

Patch diffing — the technique of analyzing the differences between a patched and unpatched binary to reverse-engineer the vulnerability being fixed — has become a primary driver of rapid post-disclosure exploitation. AI-assisted analysis has made patch diffing dramatically faster: what previously required experienced reverse engineers working for days can now be accomplished by AI agents in hours [14]. The practical consequence is that every patch release is simultaneously a vulnerability disclosure and an exploit blueprint, available to any actor capable of running an AI tool against the patch differential.

The economics of exploit development have undergone a comparable compression. Research published in Dark Reading documented AI systems generating proof-of-concept exploit code in under 15 minutes, with per-exploit costs estimated at approximately $1 — though costs vary considerably by target and context [15]. When exploit development cost drops by several orders of magnitude, the population of

actors capable of weaponizing disclosed vulnerabilities expands proportionally. Vulnerability exploitation, historically requiring the kind of expertise concentrated in nation-state offensive teams and elite criminal groups, becomes accessible to a far broader range of threat actors — state and non-state, criminal and ideological.

The defender side of this equation has not experienced comparable acceleration. Enterprise patch cycles, constrained by change management processes, compatibility testing requirements, and operational continuity concerns, have not compacted at anything approaching the rate of exploit development. Vendors take an average of 15 days to patch actively exploited vulnerabilities [13]. More troublingly, 50% of vulnerabilities listed in CISA's Known Exploited Vulnerabilities catalog — the definitive list of vulnerabilities actively abused in real attacks — remain unpatched 55 days after a fix becomes available [13]. The structural mismatch between offensive capability and defensive response is not a temporary transitional condition; it reflects deep organizational and operational constraints that cannot be resolved by individual enterprise action.

This is the fundamental asymmetry that AI-autonomous vulnerability discovery threatens to make permanent rather than transient: attackers discover and weaponize faster than defenders can patch, and AI acceleration compounds the discovery side of that equation without offering commensurate relief on the remediation side.

The Infrastructure Crisis: NVD, CVE Volume, and the Limits of Coordination

The vulnerability management ecosystem rests on a set of shared institutional resources — the CVE identifier program, NIST's National Vulnerability Database enrichment process, CVSS scoring, vendor notification channels — that were designed for a different order of magnitude of findings. The AI acceleration of vulnerability discovery is testing that infrastructure to its breaking point.

CVE submission volume provides the most direct measure of the strain. A record 48,185 CVEs were published in 2025, representing approximately a 20% year-over-year increase from approximately 39,962 published in 2024, and part of a larger trend that saw CVE submissions increase by 263% between 2020 and 2025 [5][23]. NIST's enrichment program — which adds CVSS scores, CWE classifications, and reference links to raw CVE records, transforming them into the structured data that vulnerability management tools consume — processed nearly 42,000 CVEs in 2025, 45% more than any prior year [6]. It was not enough.

In April 2026, NIST formally acknowledged the situation. The agency announced that CVE submissions during the first three months of 2026 were running nearly one-third higher than the equivalent period the prior year, and that it could no longer process submissions at the rate they arrived [6][22]. The practical consequence was a formal retreat from universal enrichment: all CVEs with an NVD publish date before March 1, 2026, were moved to a "Not Scheduled" category, meaning they will receive no enrichment from NIST [16]. Going forward, NIST will prioritize enrichment only for CVEs appearing in CISA's Known Exploited Vulnerabilities catalog, CVEs affecting federal government software, and CVEs affecting "critical software" as defined under Executive Order 14028 [6].

This policy change is consequential in ways that extend beyond NIST's operational capacity. The NVD is not merely a government database; it is the foundational data source for a broad ecosystem of vulnerability management platforms, security information and event management systems, risk scoring tools, and compliance workflows. When NVD enrichment becomes selective, the downstream effects cascade through the broad ecosystem of tools that rely on NVD as their primary structured data source. Vulnerability management programs built around CVSS scores, vendor-neutral risk quantification, and automated prioritization lose their shared reference point for the large fraction of CVEs that fall outside NIST's new priority categories.

The problem is compounded by the structure of AI-assisted discovery. When a single AI system can survey an entire codebase for under $50 and identify dozens or hundreds of distinct vulnerabilities, the disclosure pipeline faces a qualitatively different challenge than managing findings from human researchers working on single targets. A research team running AI discovery tools continuously against a large software catalog could plausibly generate more CVE-worthy findings per week than the NVD processed in earlier years — approximately 20,000 CVEs in 2021 — though actual throughput would depend heavily on codebase selection and validation rigor. The coordination infrastructure — vendor notification, triage, scoring, public disclosure — was not designed for that throughput, and there is no credible roadmap for scaling it to match.

The coordinated disclosure process is facing its own parallel strain. An analysis of approximately 44,900 AI-related projects on GitHub found that the majority lacked any support for structured security workflows or native coordinated vulnerability disclosure (CVD) tools, despite representing a rapidly expanding attack surface [17]. When researchers discover vulnerabilities in AI systems — whether AI-specific flaws or conventional bugs in AI-integrated codebases — the absence of clear disclosure channels creates friction that slows the remediation process and increases the window during which vulnerabilities remain unaddressed after discovery.

Systemic Consequences: Four Cascading Failures

The collision of AI-accelerated vulnerability discovery with a strained vulnerability management ecosystem is producing systemic failures along four distinct dimensions. These are not independent problems; they interact and amplify each other, creating a cascade that individual organizational responses alone cannot fully address.

The Deepening Patch Deficit. As vulnerability discovery volume increases — whether from AI tools operated defensively, by security researchers, or eventually by threat actors — the pool of known but unpatched vulnerabilities expands faster than enterprise remediation capacity can drain it. This is not a linear scaling problem. Each disclosed vulnerability requires triage, risk assessment, patch scheduling, deployment coordination, and validation before it can be closed. When AI systems discover not dozens but thousands of vulnerabilities per program run, the triage burden alone exceeds what most security operations teams can sustain. The result is a growing inventory of publicly known but unpatched vulnerabilities that attackers can target before organizations remediate them — and that inventory grows larger with each major AI discovery program announcement.

The operational implication is stark. When Anthropic disclosed the Claude Mythos Firefox findings — 271 zero-days in a single browser — Mozilla's security team faced the task of triaging, validating, and fixing those vulnerabilities against a disclosure timeline presumably negotiated with Anthropic's Project Glasswing under coordinated disclosure protocols. The organizational capacity required to process that volume of findings simultaneously, without interrupting ongoing development or normal security operations, represents a challenge that few organizations possess. When similar disclosure events occur at other vendors — and the trajectory of AI vulnerability research makes similar events not just possible but likely — the aggregate burden on the vendor ecosystem becomes difficult to manage.

Overload of Disclosure Pipelines. Responsible disclosure has always required a tripartite coordination among researchers, vendors, and public disclosure mechanisms. That coordination depends on each party having enough attention and organizational capacity to engage meaningfully. When the volume of findings from a single AI-assisted research program exceeds what a vendor's security team can process within standard disclosure timelines — typically 90 days from notification — the pipeline breaks down in predictable ways. Vendors may request extended timelines they are not confident they can meet. Researchers, or the AI operators they represent, may face pressure to truncate coordinated disclosure in favor of public release to protect downstream users. Either outcome degrades the system's ability to ensure that patches precede widespread exploitation.

OpenAI has noted that the traditional 90-day coordinated disclosure model was designed around the implicit assumption that the number of simultaneous disclosures from any single source would be manageable — typically measured in single or low double digits per engagement [18]. AI vulnerability

research programs operating continuously against large software catalogs break that assumption fundamentally. A new model for scaled, pipeline-oriented disclosure is needed, but no consensus architecture has yet emerged.

Asymmetric Capability Transfer. The AI systems that enable large-scale defensive vulnerability research are also available — with varying levels of restriction — to threat actors. The most capable systems, like Claude Mythos Preview, are currently restricted to vetted partners under Project Glasswing's access controls [1]. Prior-generation models, however, are widely accessible, and the research trajectory strongly suggests that today's restricted capabilities will, within months or years, be available in open or lightly restricted forms. Even current-generation models, when circumvented through jailbreaking or misuse, demonstrate concerning offensive capability.

In November 2025, a Chinese threat group used a jailbroken Claude Code instance to conduct what security researchers described as 80–90% autonomous cyber espionage operations targeting approximately 30 organizations, according to CSA's analysis of Anthropic's disclosure documentation [3]. This incident illustrated a critical asymmetry: AI capability restrictions function as a temporary barrier, not a permanent one. As AI vulnerability research capabilities diffuse — through open-weight model releases, fine-tuning on security datasets, and creative circumvention of usage restrictions — the population of actors capable of conducting AI-assisted vulnerability discovery at scale will grow substantially. The window during which AI-powered discovery provides primarily defensive benefit is likely short.

Open Source Ecosystem Fragility. A substantial fraction of the software that underpins enterprise and critical infrastructure security relies on open source codebases maintained by small teams of volunteers or lightly funded organizations. When AI discovery programs reveal large numbers of vulnerabilities in widely used open source projects, the remediation burden falls on maintainers who often lack the resources to process findings at the rate they arrive. This creates a pattern that security researchers and maintainers have identified as a failure mode distinct from individual vulnerability severity: the onslaught of well-intentioned but volume-intensive disclosure can produce burnout, create noise that masks genuinely critical findings, and generate plausible-looking reports that waste human review time without improving actual security posture [17]. When the disclosure volume from a single AI-assisted research program exceeds what an open source project's maintainers can absorb, the coordination process that responsible disclosure depends on breaks down regardless of individual researcher intent.

Enterprise Security in a Post-Threshold World

The systemic consequences described above do not distribute evenly across the enterprise landscape. Organizations with mature vulnerability management programs, well-resourced security teams, and strong vendor relationships will face severe strain; organizations operating with minimal security investment may find that the traditional risk calculus — patch slowly, prioritize only critical findings — is no longer viable against a threat landscape where AI-assisted exploitation makes low-priority vulnerabilities newly dangerous.

For enterprise security programs, the most immediate consequence of the collapsing exploit window is the obsolescence of patch-cycle-based prioritization as a primary risk management strategy. When 28.3% of exploited vulnerabilities are weaponized within 24 hours of disclosure [4], a monthly patch cycle provides meaningful protection only against the fraction of vulnerabilities that threat actors have not yet targeted. This does not mean that patching is unimportant — it means that the framing of patch management as a risk reduction activity, rather than a continuous operational requirement, no longer matches the threat environment.

The enterprise response requires a fundamental reorientation in how vulnerability risk is understood and managed. Exposure management — understanding which vulnerabilities are present in an environment and whether those environments are reachable by attackers — must become a continuous, near-real-time capability rather than a periodic audit function. Network segmentation, application layer controls, and behavioral detection capabilities become compensating controls for the inevitable inventory of known vulnerabilities that an organization cannot patch at the rate they are being disclosed. The goal of eliminating known vulnerabilities from the environment, always aspirational, becomes explicitly secondary to limiting the blast radius when unpatched vulnerabilities are exploited.

The AI acceleration of vulnerability discovery also suggests that enterprises should approach their own software inventory differently. The finding that years-old vulnerabilities in widely audited codebases — OpenBSD, FreeBSD, Firefox — remained undetected until AI analysis surfaced them implies that the security posture of legacy and long-stable systems cannot be assumed from the absence of prior findings. Software that has "passed" security review under human-speed analysis may harbor vulnerabilities invisible to that analysis but detectable by AI systems. This argues for a proactive posture: enterprises running AI-assisted security scanning against their own critical systems, rather than waiting for a vendor or external researcher to discover and disclose findings, gain the defensive benefit of AI-accelerated discovery before the same tools are applied offensively.

Threat intelligence programs must similarly adapt to a changed signal environment. When vulnerability weaponization timelines compress to hours, the traditional cycle of monitoring threat feeds and generating patch priority recommendations becomes too slow to provide meaningful operational

guidance. Real-time signals — darkweb exploit market activity, proof-of-concept publication, active scan traffic for specific vulnerability signatures — must feed into automated triage and response workflows that can initiate remediation actions without waiting for human review of each finding.

The vendor and supply chain dimension of enterprise risk requires particular attention. The concentration of software supply chains around a small number of widely used open source components means that a major AI-assisted vulnerability discovery program targeting common libraries — OpenSSL, glibc, the Linux kernel — could simultaneously expand the attack surface for thousands of downstream organizations. Enterprises with strong software bill of materials (SBOM) capabilities will be positioned to identify affected components quickly; those without SBOM visibility will face extended exposure as they attempt to inventory dependencies manually.

Conclusions and Recommendations

The convergence of AI-autonomous vulnerability discovery, compressed exploitation timelines, and failing vulnerability coordination infrastructure represents a structural transition in the threat environment, not a temporary perturbation. The security institutions — responsible disclosure programs, vulnerability databases, patch cadence models — that have governed the patch-and-exploit cycle for the past two decades were designed for a world where discovery was slow, expensive, and concentrated in small expert communities. That world no longer exists.

Several strategic adjustments are essential for organizations operating in this environment.

Move from periodic patch management to continuous exposure management. The patch cycle model implicitly assumes that vulnerability discovery and exploitation both occur on timescales that human-paced processes can track. Neither assumption holds. Effective risk reduction now requires continuous inventory of vulnerable components, real-time integration of threat intelligence into prioritization decisions, and automated response workflows capable of acting within hours of high-confidence exploitation signals — not days or weeks.

Proactively adopt AI-assisted security scanning as a defensive capability. Organizations that wait for AI-discovered vulnerabilities to be disclosed and patched through normal channels will experience extended exposure windows, particularly for legacy or widely-deployed internal systems. Enterprises with the operational capability should pursue access to AI-assisted security scanning tools — including programs like Project Glasswing for qualifying organizations — and integrate AI-assisted code analysis into development pipelines. Discovering vulnerabilities through a defensive program is strongly preferable to learning of them through a breach notification.

Invest in compensating controls and blast radius limitation. When AI-accelerated discovery ensures that some fraction of vulnerabilities will be exploited before patches are available, the strategic priority becomes limiting damage when exploitation occurs. Network micro-segmentation, application-layer controls that constrain what an exploited component can access, behavioral detection capabilities that identify post-exploitation activity, and robust incident response programs are no longer optional investments for organizations with critical data or systems.

Engage the policy and coordination process. The NVD enrichment crisis, the breakdown of responsible disclosure at scale, and the asymmetric diffusion of AI offensive capabilities are structural problems that no individual organization can solve unilaterally. Security leaders should actively participate in policy processes — CISA stakeholder engagement, NIST framework development, vendor disclosure coordination councils — that will determine how the security ecosystem adapts. The decisions made over the next 18 months about how AI vulnerability research is governed, how disclosure timelines are adjusted, and how national vulnerability infrastructure is resourced will shape the operating environment for years.

Establish organizational policies for AI-assisted security tooling. The same AI capabilities that enable defensive vulnerability research can be misused — deliberately or through negligence — in ways that create new risk. Organizations should establish clear governance frameworks covering which AI-assisted security tools may be used, under what authorization, against what targets, and with what disclosure obligations for findings. Absence of governance creates liability exposure and increases the likelihood of internal misuse or inadvertent harm.

Support open source security capacity. A large fraction of enterprise security ultimately rests on the security of open source components maintained with limited resources. Enterprises that depend heavily on open source software have a stake in ensuring that the maintainer communities for those projects have the capacity to process AI-generated vulnerability findings responsibly. Contributing resources — financially, through dedicated security engineering time, or through participation in coordinated disclosure programs — to open source security programs is both a direct risk reduction measure and an investment in the shared security infrastructure on which the entire ecosystem depends.

CSA Resource Alignment

The threat landscape described in this paper intersects directly with the frameworks and research programs that the Cloud Security Alliance has developed to address AI and cloud security challenges.

The **MAESTRO (Multilayer Epistemic Architecture for Securing Transformative Reasoning Operations)** framework addresses the threat modeling of agentic AI systems, including the multi-stage attack lifecycles that characterize AI-assisted vulnerability exploitation. The autonomous vulnerability research and exploitation chains demonstrated by systems like Mythos Preview represent exactly the class of multi-step agentic capability that MAESTRO's threat modeling taxonomy was designed to characterize. Security architects evaluating AI-assisted security tools or assessing exposure to AI-powered offensive capabilities should apply MAESTRO's analysis of agentic control points and failure modes.

The **AI Controls Matrix (AICM)** provides the governance structure within which AI tool deployment — including AI-assisted vulnerability research tools — should be assessed. The AICM's controls for model access governance, output validation, and human oversight are directly applicable to the governance questions raised by AI vulnerability research programs: who may operate these tools, against what targets, with what validation of findings before action, and with what disclosure obligations.

The **STAR (Security Trust Assurance and Risk) for AI** program offers a structured pathway for organizations to assess their security posture relative to AI-specific risks, including the operational dependencies on AI systems in security tooling. As AI-assisted vulnerability scanning becomes a standard security operations capability, STAR for AI provides a mechanism for assessing and benchmarking that capability in a consistent, auditable framework.

CSA's **Using AI for Offensive Security** report (2024) [19] established the foundational analysis of how large language models and AI agents could be applied across the offensive security lifecycle. The developments described in this paper represent the next phase of that trajectory: the transition from AI as an assistive tool in human-directed security research to AI as the primary locus of discovery and exploitation capability, with humans in an oversight and governance role rather than an operational one. Organizations building on the 2024 report's implementation guidance should treat the Mythos-class capability threshold as a prompt to revisit their defensive assumptions.

CSA's **Zero Trust** guidance provides the architectural principle most directly applicable to a world where known vulnerabilities cannot be reliably patched before exploitation: assume breach, minimize lateral movement, enforce least-privilege access at every boundary. Zero Trust does not solve the patch deficit problem, but it substantially limits the blast radius when exploitation of unpatched vulnerabilities occurs — converting what might be a catastrophic breach into a contained incident.

References

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

[2] Cybersecurity News. "Claude Mythos AI Model Uncovers 271 Zero-Day Vulnerabilities in Firefox." Cybersecurity News, April 2026.

[3] Cloud Security Alliance AI Safety Initiative. "Claude Mythos and the AI Autonomous Offensive Threshold." CSA Labs, April 2026.

[4] Security Boulevard. "46 Vulnerability Statistics 2026: Key Trends in Discovery, Exploitation, and Risk." Security Boulevard, March 2026.

[5] The Hacker News. "NIST Limits CVE Enrichment After 263% Surge in Vulnerability Submissions." The Hacker News, April 2026.

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

[7] Google Project Zero. "From Naptime to Big Sleep: Using Large Language Models To Catch Vulnerabilities In Real-World Code." Project Zero Blog, October 2024.

[8] Google Cloud. "Defending Your Enterprise When AI Models Can Find Vulnerabilities Faster Than Ever." Google Cloud Blog, 2026.

[9] Infosecurity Magazine. "Google OSS-Fuzz Harnesses AI to Expose 26 Security Vulnerabilities." Infosecurity Magazine, 2024.

[10] Trend Micro. "Introducing ÆSIR: Finding Zero-Day Vulnerabilities at the Speed of AI." Trend Micro Research, January 2026.

[11] Fang, Richard, et al. "LLM Agents can Autonomously Exploit One-day Vulnerabilities." arXiv:2404.08144, April 2024.

[12] Fang, Richard, et al. "Teams of LLM Agents can Exploit Zero-Day Vulnerabilities." arXiv:2406.01637, June 2024.

[13] Hive Security. "From CVE to RCE in Hours: The Collapse of the Exploitation Window." Hive Security Blog, 2026.

[14] CERT-EU. "AI is changing the economics of vulnerability discovery. Defenders should adapt now." CERT-EU Blog, 2026.

[15] Dark Reading. "PoC Code in 15 Minutes? AI Turbocharges Exploitation." Dark Reading, 2025.

[16] Infosecurity Magazine. "NIST Drops NVD Enrichment for Pre-March 2026 Vulnerabilities." Infosecurity Magazine, April 2026.

[17] Carnegie Mellon University Software Engineering Institute. "Protecting AI from the Outside In: The Case for Coordinated Vulnerability Disclosure." SEI Blog, 2025.

[18] OpenAI. "Scaling Security with Responsible Disclosure." OpenAI, 2025.

[19] Lundqvist, Adam, and Kirti Chopra. "Using AI for Offensive Security." Cloud Security Alliance AI Technology and Risk Working Group, 2024.

[20] The Hacker News. "Anthropic's Claude Mythos Finds Thousands of Zero-Day Flaws Across Major Systems." The Hacker News, April 2026.

[21] Help Net Security. "Anthropic's New AI Model Finds and Exploits Zero-Days Across Every Major OS and Browser." Help Net Security, April 2026.

[22] Help Net Security. "NIST Admits Defeat on NVD Backlog, Will Enrich Only Highest-Risk CVEs Going Forward." Help Net Security, April 2026.

[23] Hive Pro. "The CVE Deluge of 2025: Why It's More Than Just a Number." Hive Pro Blog, 2025.

[24] Deep Strike. "Vulnerability Statistics 2025: CVE Surge & Exploit Speed." Deep Strike, 2025.