CSAI Foundation

Cloud Security Alliance AI Safety Initiative

When the Model Becomes the Red Team

Threat Model and Governance for AI-Autonomous Vulnerability Discovery

Unofficial AI-assisted Research

2026-04-12

Table of Contents

Executive Summary

The security industry has long relied on a foundational assumption: discovering vulnerabilities requires deep human expertise, significant time investment, and specialized knowledge that limits the pool of capable adversaries. That assumption is no longer valid. In 2024 and 2025, a series of landmark demonstrations established that large language models (LLMs) can autonomously discover and reproduce previously unknown, exploitable vulnerabilities in widely deployed software — including bugs that persisted undetected for decades despite intensive manual review. By early 2026, Anthropic's research team had validated more than 500 high-severity zero-day vulnerabilities using Claude Opus 4.6 [1], and its successor model, Claude Mythos Preview, autonomously identified thousands of additional flaws across every major operating system and web browser before restricted release [2].

These developments are not theoretical. They represent a structural inflection point in the vulnerability lifecycle: a world in which the time required to discover and weaponize a software flaw is compressing rapidly for threat actors with access to capable AI tooling — a category that is expanding as frontier model capabilities diffuse — while the time required to patch and deploy fixes remains measured in weeks and months. The gap between these two timelines — already a persistent structural challenge — is now widening at a pace that current patch deployment cycles cannot match.

This whitepaper provides security leaders with a grounded threat model for AI-autonomous vulnerability discovery, examines the governance and disclosure frameworks that are straining to adapt, and offers practical guidance for organizations that must now defend against an adversary whose reconnaissance and exploitation capabilities have been fundamentally transformed. It also connects these findings to existing CSA frameworks, particularly MAESTRO for agentic AI threat modeling and the AI Controls Matrix (AICM), which together provide the foundational vocabulary for reasoning about agentic offensive threats.

The core finding is both simple and consequential: organizations must treat AI-autonomous vulnerability discovery not as a future risk to be monitored but as a present capability that is already reshaping the threat landscape. Defenders whose governance and patch timelines remain calibrated for pre-AI norms face growing exposure, and the investment required to adapt those norms is available today.

Introduction and Background

The Long Road to Autonomous Discovery

Automated vulnerability discovery is not a new idea. For decades, the security community has employed static analysis tools, dynamic fuzzers, and symbolic execution engines to augment human-led code review. These tools have delivered real value — OSS-Fuzz, Google's continuous fuzzing infrastructure, has identified more than 13,000 vulnerabilities and 50,000 bugs across more than 1,000 open-source projects since its launch [14]. Yet automated tools have historically operated within sharp constraints: they excel at coverage-driven input generation but struggle with the semantic understanding required to recognize subtle logical flaws, incomplete bound checks, or algorithm-level behavioral anomalies that an experienced human researcher would immediately identify as suspicious.

Large language models change this picture by bringing something automated tools have historically lacked: the ability to reason about code in context. Unlike fuzzers, which treat code as a black box to be probed with random or mutation-based inputs, LLMs can process commit histories, generate plausible hypotheses about the intent behind code changes, recognize unsafe patterns across different functions, and identify locations where a recently patched vulnerability might have siblings that were left unaddressed. This represents a qualitative shift rather than a quantitative one. The question facing security teams is not whether AI can be faster or cheaper than human researchers — it is whether AI fundamentally changes the economics, accessibility, and scale of offensive vulnerability research.

The evidence from 2024 through early 2026 suggests the answer is yes.

From Experimental to Demonstrated

Google Project Zero and Google DeepMind's Big Sleep agent, described in detail in a Project Zero blog post from November 2024 [3], represents the first publicly documented case of an AI agent discovering a previously unknown exploitable memory-safety vulnerability in widely used production software. The target was SQLite, an embedded database engine present in hundreds of millions of devices. Working from variant analysis — beginning with a recently patched vulnerability and asking the model to identify related unpatched code paths — the agent discovered a stack buffer underflow in the `seriesBestIndex` function. The vulnerability arose from improper handling of a sentinel value used to indicate ROWID constraints, causing index calculations to produce negative values that wrote beyond buffer boundaries. The reproduction case was elegantly simple: a single SQL query that any developer could execute. Crucially, 150 CPU-hours of AFL fuzzing had failed to find the same flaw, largely because the OSS-Fuzz harness lacked the relevant extension and alternative configurations were running on outdated code.

The Big Sleep finding was a proof of concept at scale. It demonstrated that an AI agent could perform the kind of semantic reasoning that security researchers associate with expert human intuition — analyzing commit diffs, hypothesizing about unsafe pattern variants, adapting when initial test approaches failed, and producing a coherent root-cause analysis. The significance was recognized immediately: this was not a tool finding the same class of bug a fuzzer would eventually reach; it was an agent producing outputs that functionally matched what a senior researcher would develop, only faster and without fatigue.

Building on that foundation, Anthropic's Frontier Red Team in February 2026 deployed Claude Opus 4.6 against a broad range of production open-source codebases [1]. The setup was deliberately minimal: a virtual machine, access to the target projects, standard utilities, and common analysis tools including debuggers and fuzzers, but no special instructions on how to use them. Over the course of the research campaign, Claude identified, reproduced, and validated more than 500 high-severity vulnerabilities. The examples are instructive. In GhostScript, Claude analyzed commit history to identify incomplete stack bounds checking, then located a similar unpatched code path triggered by malformed PostScript processing. In OpenSC, it detected unsafe successive string concatenation operations capable of causing buffer overflows when processing smart card data. In CGIF, it recognized how LZW compression's dictionary reset mechanism could cause output to exceed buffer capacity — a finding that required understanding algorithm behavior at a conceptual level, not merely pattern matching against known-bad idioms.

Less than two months later, Anthropic announced Claude Mythos Preview and Project Glasswing [2], marking a further capability step. Claude Mythos demonstrated 83.1 percent performance on the CyberGym cybersecurity vulnerability reproduction benchmark, compared to 66.6 percent for Claude Opus 4.6 [2]. The model autonomously identified thousands of previously unknown vulnerabilities across every major operating system and every major web browser, including a 27-year-old vulnerability in OpenBSD's TCP SACK implementation and a 16-year-old flaw in FFmpeg's H.264 codec that had survived decades of expert human review and continuous automated testing [12]. Because of the severity of these capabilities, Anthropic restricted access to Claude Mythos to a limited group of industry partners including AWS, Apple, Broadcom, Cisco, CrowdStrike, Google, JPMorgan Chase, the Linux Foundation, Microsoft, NVIDIA, and Palo Alto Networks under a structured responsible-use framework.

These are not isolated experiments. They are waypoints on a trajectory of rapid capability improvement, though whether this pace is sustained will depend on factors including training data saturation and the development of new evaluation categories.

The Capability Landscape

What AI Models Can Do Today

AI-autonomous vulnerability discovery now encompasses several distinct capability categories, each with different implications for threat modeling. Understanding these categories is essential to calibrating the appropriate defensive response.

The first and most mature capability is variant analysis: given a recently patched or publicly disclosed vulnerability, an AI agent can analyze the relevant code and its dependencies to identify related flaws that share the same root cause but were not fixed in the original patch. This is how Big Sleep found the SQLite vulnerability, and it represents a significant accelerant to the post-disclosure exploitation window. When a vendor patches a vulnerability, they typically disclose enough technical detail for downstream users to understand what changed. AI systems can use exactly that detail to locate unpatched variants in other codebases or in the same codebase under different execution paths.

The second capability is direct static analysis at scale. LLMs can read source code and identify patterns associated with memory corruption, improper authentication, injection flaws, and other vulnerability classes with a fluency that resembles human expertise. Unlike traditional static analysis tools, which operate on syntactic patterns and predefined rules, LLMs understand semantic context — they can reason about what a function is supposed to do, identify where it deviates from correct behavior, and explain the conditions under which deviation becomes exploitable. This capability is demonstrated in the OpenSC and CGIF findings from Anthropic's research: neither vulnerability is the kind that a rule-based scanner would reliably detect.

The third capability is end-to-end exploit reproduction. CVE-Genie, a multi-agent framework described in a September 2025 arxiv preprint [4], demonstrated that AI systems can not only identify vulnerabilities but reproduce existing ones end-to-end from CVE entries. Given a CVE description as input, the system gathers relevant resources, reconstructs the vulnerable environment, and produces a verifiable exploit. Running against 841 CVEs published between June 2024 and May 2025, CVE-Genie successfully reproduced approximately 51 percent — 428 CVEs across 267 projects, 141 CWEs, and 22 programming languages — at an average cost of $2.77 per CVE. This is not vulnerability discovery; it is vulnerability operationalization at commodity pricing.

The fourth capability is autonomous penetration testing. PentestGPT, originally described at USENIX Security 2024 [5] and substantially extended into a fully autonomous agent by late 2025, achieved an 86.5 percent success rate on the XBOW validation benchmark across 104 challenges — a result attributable to the extended autonomous agent version. While fully autonomous end-to-end pentesting against complex, real-

world environments remains constrained — published research benchmarks report success rates ranging from 21 to 31 percent for fully autonomous configurations against realistic scenarios across multiple evaluations — the trajectory of improvement has been consistent.

Together, these capabilities define a threat surface that is qualitatively different from anything security teams have previously had to account for.

The Industrialization of Offensive Research

What makes these capabilities especially significant is not their technical sophistication in isolation but what they imply for the economics and accessibility of offensive research. Historically, the discovery of a high-severity zero-day required months of focused effort from a skilled researcher who combined deep domain expertise with specialized tooling. That expertise was rare, expensive, and difficult to acquire — which is why nation-state threat actors and well-funded criminal organizations were the primary market for zero-day vulnerabilities, and why zero-day brokers could command prices measured in hundreds of thousands to millions of dollars.

AI systems dissolve these economic barriers. When CVE-Genie can reproduce known exploits at $2.77 per attempt, and when Claude Opus 4.6 can discover novel high-severity vulnerabilities using nothing more than a VM and standard tools, the cost curve for offensive capability shifts dramatically. Capabilities that once required a team of expert researchers over months can now be approximated by an AI-augmented actor over days, or in some cases hours — Anthropic's red team documented a FreeBSD remote root exploit written autonomously in four hours [1].

This industrialization effect is not hypothetical. CrowdStrike reported an 89 percent year-over-year increase in AI-enabled adversary attacks in 2025 [6], and the combination of AI-powered reconnaissance, vulnerability identification, and exploit generation is increasingly visible in threat intelligence from multiple vendors. The implication for security teams is that the threat actor population capable of conducting sophisticated offensive operations is expanding, and the skill floor required to participate is dropping.

Threat Model

Actors, Objectives, and Attack Vectors

A rigorous threat model for AI-autonomous vulnerability discovery must account for the full range of actors who may benefit from these capabilities, the objectives they are likely to pursue, and the attack vectors that AI capabilities make newly accessible or substantially easier.

Nation-state threat actors represent the most capable tier. These organizations have long invested in zero-day research and already employ teams of skilled vulnerability researchers. For them, AI augments an existing capability: it accelerates the discovery of variants after initial patch analysis, enables parallel research across multiple target codebases simultaneously, and reduces the time from discovery to weaponized exploit. The practical effect is that nation-state actors can maintain larger portfolios of active zero-days and respond more quickly to patch releases by identifying unpatched siblings before defenders deploy fixes. CrowdStrike's 2026 Global Threat Report records the mean eCrime breakout time — the interval between initial access and lateral movement — at 29 minutes in 2025, with the fastest recorded breakout at 27 seconds [6], reflecting an acceleration in operational tempo that compounds the advantages conferred by AI-assisted reconnaissance and exploitation.

Criminal threat actors represent the second tier, and it is here that the industrialization effect is most consequential. Financially motivated attackers have historically relied on publicly disclosed vulnerabilities and commodity exploit kits because the cost of original zero-day research was prohibitive relative to expected returns. AI systems lower that cost threshold substantially. An organized criminal group with moderate technical resources can now operate vulnerability research pipelines that would previously have required sophisticated offensive security capabilities. The ransomware and extortion ecosystem, which has already demonstrated significant organizational sophistication, gains a meaningful capability upgrade from AI-assisted offensive research.

A third actor category — security researchers operating under dual-use conditions — must be considered not as a threat in itself but as a source of capability proliferation. The same tools that enable Project Glasswing partners to find and fix vulnerabilities before they are exploited are, in most respects, the same tools an adversary would use to find and exploit those vulnerabilities before they are patched. The difference lies in intent, authorization, and the governance framework surrounding access — not in the underlying technology. This symmetry creates genuine policy challenges.

The Discovery-Exploitation Gap

The most consequential structural change in the threat landscape is the collapse of the time gap between vulnerability discovery and exploitation. Security operations have historically been calibrated around an assumption that most vulnerabilities are exploited days to weeks after public disclosure, giving defenders time to assess, prioritize, and patch before attackers arrive. That assumption was already under pressure before AI; it is now dangerously optimistic.

Research published across multiple sources documents a consistent pattern. The median time to exploit following public disclosure dropped from approximately 32 days to roughly five days by 2025 [13, 7]. Organizations' median time to patch critical vulnerabilities, by contrast, runs approximately 32 to 55 days depending on vulnerability category — with critical items in CISA's Known Exploited Vulnerabilities catalog taking an average of 55 days to reach 50 percent remediation [7] — creating a structural exposure window of

three to seven weeks during which organizations are vulnerable to exploits targeting disclosed flaws. The asymmetry between these timelines is not a temporary operational gap; it is a structural characteristic of the current threat environment.

AI-autonomous vulnerability discovery worsens this asymmetry at both ends. On the offensive side, AI systems can convert a patch analysis into a weaponized exploit faster than any human team. On the defensive side, AI-assisted variant discovery actually presents an opportunity — but only for organizations that have access to such capabilities and can coordinate disclosures effectively. For the majority of organizations that are consumers of patches rather than producers, the practical effect of AI-accelerated exploitation is that patches must be deployed far more quickly than current operational rhythms support.

This is not merely a technical problem. It is an organizational and governance problem. Patch deployment timelines are constrained by testing requirements, change management processes, business continuity concerns, and in many cases the need to coordinate across complex supply chains. AI cannot solve these organizational constraints, which means that even as offensive capability accelerates, the defensive response function remains bounded by the pace of human institutional processes.

The Asymmetry Problem

The asymmetry between offense and defense in AI-assisted vulnerability research is real and structural, but it is important to characterize it accurately to avoid paralysis or fatalism. The asymmetry is not that defenders have no AI capabilities — defenders are increasingly deploying AI for behavioral detection, threat hunting, and automated response. The asymmetry is more specific: offense requires discovering one exploitable path, while defense requires securing all of them. This "one to many" structural problem is not new, but AI amplifies it. An AI system conducting offensive research can pursue thousands of potential vulnerability hypotheses in parallel, while a defense team must prioritize patches across a finite resource pool. The attacker who finds a single exploitable vulnerability gains an initial foothold; the defender's goal is not to eliminate every vulnerability but to ensure that exploitation does not yield mission success — through detection, segmentation, and compensating controls that interrupt the kill chain even when initial compromise occurs.

A further asymmetry concerns visibility. Defenders frequently lack clear visibility into the attack surface they are responsible for protecting. Dependency chains in modern software are long and often poorly documented; open-source components embedded in commercial products may not receive patches promptly even when upstream fixes are available; and the AI training or deployment infrastructure of organizations deploying AI systems itself introduces new attack surfaces that are difficult to enumerate and assess. An AI system conducting offensive research can reason about these dependency relationships from publicly available information, while defenders may not have equivalent internal visibility into their own systems.

Governance and the Disclosure Framework

Responsible Disclosure Under Pressure

The responsible disclosure ecosystem that has evolved over the past two decades — built around 90-day timelines, coordinated patching, and vendor engagement processes — was designed for a world where vulnerabilities are discovered by skilled human researchers who can make nuanced judgments about severity, exploitability, and disclosure timing. AI-autonomous discovery challenges this model in ways that are only beginning to be addressed.

When an AI system identifies 500 vulnerabilities in a matter of weeks, the downstream coordination requirements are entirely different from anything the disclosure ecosystem was built to handle. Anthropic has published a coordinated vulnerability disclosure policy [8] specifically addressing AI-discovered vulnerabilities. The policy maintains a 90-day standard timeline, with a 7-day compressed timeline for actively exploited or critical vulnerabilities. It requires human security researcher review and confirmation before any submission, explicitly labels AI-discovered findings as such, and commits to simultaneous notification of all affected maintainers when vulnerabilities affect multiple projects. For non-responding maintainers, the policy specifies escalation to external vulnerability coordinators at 30 days.

This framework represents a thoughtful adaptation of existing disclosure norms, but it is worth examining its implications carefully. The requirement for human review before disclosure is a meaningful safeguard against false positives and against overwhelming maintainers with unvalidated machine-generated findings. Yet it also represents a bottleneck: if AI systems can discover vulnerabilities faster than human researchers can validate them, the practical effect may be that validated disclosures represent only a fraction of what has actually been found. The validated backlog that is not yet disclosed creates its own risk profile — a portfolio of known vulnerabilities that defenders cannot yet remediate because they have not been notified.

The 90-day disclosure timeline also warrants scrutiny. This standard was established at a time when the exploit development cycle following disclosure was measured in days to weeks. When AI can compress that development cycle to hours, a 90-day window during which technical details are publicly available but patches may not yet be deployed may represent an extended period of elevated risk rather than a sufficient buffer for remediation. It is also worth noting the counterargument: compressing disclosure timelines further risks pushing organizations to deploy incomplete patches under time pressure, potentially introducing new vulnerabilities in the remediation itself. The gap between when Anthropic publicly knew about vulnerabilities and when patch details were shared with the affected maintainers is a policy question with real security implications — one where the community must balance the risks of prolonged exposure against the risks of rushed remediation.

The Glasswing Precedent and Access Governance

Project Glasswing [2] represents a novel governance model worth examining in detail, because it addresses a genuine dilemma: what do you do with an AI system that is simultaneously the most effective tool for finding and fixing vulnerabilities, and potentially one of the most dangerous systems ever built for offensive security purposes?

Anthropic's response was to restrict initial access to a vetted group of organizations with direct responsibility for critical infrastructure, commit $100 million in model usage credits to support their defensive work, and establish a 90-day reporting requirement for participants to share findings and best practices. The participating organizations — spanning cloud infrastructure, consumer hardware, financial services, open-source foundations, and cybersecurity vendors — represent a broad coalition of defenders who collectively bear responsibility for systems used by billions of people.

The precedent this sets is significant. It suggests that certain AI capabilities may be inappropriate for general availability not because they are inherently harmful but because the risks of misuse outweigh the benefits of broad access until defensive ecosystems have had time to mature. This is a recognizable principle in other dual-use technology contexts, but it is relatively new territory for AI deployment. The Glasswing model implies that responsible deployment of frontier offensive security AI requires governance structures that go beyond standard terms of service or usage policies — it requires active partnerships with defenders, structured reporting obligations, and a clear theory of how controlled access translates into improved security outcomes before broader availability.

Not every organization developing AI systems with offensive security capabilities will adopt this approach voluntarily. The governance question of how the broader industry coordinates around AI models that are capable of autonomous vulnerability discovery — including what access controls, usage policies, and reporting obligations should apply — remains unresolved and urgent.

The Disclosure Norm Vacuum

Beyond individual vendor policies, the security community faces a structural gap in norms for AI-assisted mass vulnerability discovery. The current coordinated disclosure ecosystem functions through a network of bilateral relationships between security researchers, vendors, and coordination bodies such as CERT/CC and CISA's vulnerability disclosure program. These relationships were built for a world of individual or small-team research, where the volume of findings from any single source was bounded by human capacity.

When an AI system identifies hundreds of vulnerabilities in a matter of days, the existing coordination infrastructure — built around email chains, ticketing systems, and relationship-based prioritization — is not designed to process that volume without significant strain. The risk is not merely administrative: maintainers who are overwhelmed by AI-generated disclosure reports may deprioritize genuine high-severity findings,

miss critical notifications in a flood of lower-priority ones, or fail to coordinate effectively across dependent projects. The patch quality that emerges from rushed remediation under volume pressure may itself introduce new vulnerabilities.

Addressing this gap will require deliberate investment in disclosure infrastructure — automated triage mechanisms, standardized machine-readable formats for AI-generated vulnerability reports, and clearer expectations about volume management between AI-assisted research teams and the projects they target. It will also require expanding the capacity of intermediary coordination bodies, which currently lack the staffing and tooling to serve as effective brokers for high-volume AI-generated disclosures.

Defensive Posture and Organizational Guidance

Rethinking Vulnerability Management Timelines

The most immediate practical implication of AI-accelerated exploitation is that patch deployment timelines must be compressed for critical vulnerabilities. Organizations that are currently operating with 30- to 60-day patch cycles for high-severity vulnerabilities are carrying a risk exposure that was already elevated before AI; it is now untenable for the highest-risk systems and software. Security leadership must engage with change management, IT operations, and business leadership to establish tiered patch deployment targets that reflect the actual exploitation timeline AI enables.

For actively exploited vulnerabilities and for critical findings with public proof-of-concept code, a 24- to 48-hour patch deployment window for internet-facing systems is appropriate for most organizations. For high-severity vulnerabilities without active exploitation, a 7-day target is reasonable, with 14 days as an outer bound. These recommendations are more aggressive than current framework defaults — CISA's Known Exploited Vulnerabilities catalog sets a 14-day remediation target for actively exploited findings, and NIST SP 800-40 provides broader patching lifecycle guidance calibrated to pre-AI exploitation timescales. The departure is intentional: AI-assisted exploitation has compressed the window between public disclosure and weaponized attack to a degree those frameworks did not anticipate, and organizations should treat these targets as operational guidance appropriate to the current threat environment rather than as general replacements for existing framework requirements. Achieving these timelines requires organizational infrastructure that many enterprises have not yet built: pre-authorized emergency change processes, tested rollback procedures, staging environments that can validate patches rapidly, and on-call capacity to execute emergency deployments.

Adopting AI-Assisted Defensive Research

The same capabilities that make AI dangerous as an offensive tool make it valuable as a defensive one. Organizations responsible for significant software — whether as developers of widely used libraries, maintainers of open-source projects, or operators of custom-developed applications — should evaluate AI-assisted security review as a complement to existing secure development lifecycle (SDLC) processes. Running AI-based vulnerability analysis against code before release, particularly against high-risk components such as memory management, authentication logic, and cryptographic implementations, can surface issues that automated static analysis and manual review miss.

The key operational insight from Anthropic's and Google's research is that AI vulnerability discovery works best when targeted: focusing on recent changes, code paths adjacent to previously patched vulnerabilities, and components with complex algorithmic behaviors. This targeted approach produces higher-quality findings than open-ended scanning and is more tractable for human validation pipelines. Organizations incorporating AI into their security testing programs should design these focused workflows rather than treating AI as a general-purpose scanner.

It is equally important to establish clear processes for handling AI-generated findings. The CVE-Genie research demonstrates that AI systems can produce verified findings at scale, but verification of exploitability remains a resource-intensive human task. Organizations should not publish or remediate AI-generated security findings without human review, both to avoid false positives and to ensure that the contextual judgment required for severity assessment — considering business impact, exploitability constraints, and mitigating controls — is preserved.

Architectural Controls and Exposure Reduction

Because AI-assisted exploitation compresses the time between discovery and weaponized attack, reducing exposure — the number of systems and surfaces that can be reached and attacked — becomes more valuable than it has historically been. Several architectural approaches directly reduce exposure to AI-accelerated exploitation.

Memory-safe programming languages eliminate entire classes of vulnerability that AI systems are particularly effective at finding. The migration of critical system components from C and C++ to Rust, Go, or other memory-safe alternatives is a long-term investment that directly reduces the attack surface available to memory-corruption-focused AI agents like Big Sleep and Claude Opus 4.6. Organizations that contribute to or depend on open-source software written in memory-unsafe languages should evaluate migration paths for the highest-risk components, even when full migration is not feasible in the near term.

Software composition analysis (SCA) and software bill of materials (SBOM) programs become more critical when AI-assisted variant analysis can rapidly identify vulnerable patterns across dependency trees. Organizations that maintain comprehensive, up-to-date SBOMs are better positioned to respond quickly when AI-discovered vulnerabilities surface in their dependencies — they can identify affected systems, prioritize remediation, and implement temporary mitigations more quickly than organizations without this visibility. SBOM-based alerting, integrated with vendor vulnerability feeds and national vulnerability databases, can substantially compress the time between external disclosure and internal awareness.

Network segmentation and zero trust architecture reduce the blast radius when vulnerabilities are exploited before patches are deployed. CrowdStrike's 2026 research documents a mean eCrime breakout time of 29 minutes, with the fastest recorded instance at 27 seconds [6] — timescales at which traditional detection-and-response workflows cannot reliably interrupt lateral movement in the most aggressive intrusions. The segmentation must exist architecturally, not merely as a detection trigger.

Threat Intelligence and Detection Engineering

The acceleration of offensive AI capability creates specific opportunities for detection engineering that security teams should exploit. AI-assisted attacks often exhibit detectable patterns — systematic enumeration, rapid variant payloads, and tool usage signatures that reflect underlying automation — though more sophisticated implementations may deliberately pace their activity to blend with human-generated patterns. Training detection models on AI-assisted attack signatures — rather than only on human-generated attack patterns — is increasingly important as these tools become more widely deployed.

Threat intelligence programs should track the emerging landscape of AI-assisted offensive tools, including not only commercial and state-sponsored capabilities but open-source frameworks like PentestGPT that are accessible to a broad range of actors. Understanding which tools are being used by which threat groups, and what their characteristic signatures look like at the network and endpoint level, allows detection teams to tune rules and models appropriately. The MITRE ATT&CK framework provides a baseline taxonomy, but it has not yet fully incorporated the agent-specific attack patterns that emerge from AI-assisted operations — detection teams should supplement it with AI-specific threat intelligence.

Workforce and Capability Development

The implications of AI-autonomous vulnerability discovery for the security workforce are significant and deserve deliberate attention. The skill set required to evaluate AI-generated vulnerability findings — understanding model outputs, validating exploit proofs of concept, assessing false positive rates, and making informed severity judgments — is distinct from traditional manual vulnerability research skills. As AI becomes a routine component of both offensive and defensive security research, the security workforce must develop literacy in AI system capabilities and limitations.

Organizations should invest in training for security engineers on AI-assisted vulnerability assessment methodologies, including practical experience with AI-driven analysis tools. Equally important is ensuring that security leadership understands the strategic implications of AI in the threat landscape — the decisions about patch timelines, SBOM investment, memory-safe migration, and disclosure participation are governance decisions, not purely technical ones, and they require informed executive sponsorship.

CSA Resource Alignment

The threat model and recommendations in this paper connect directly to several foundational CSA frameworks that provide complementary guidance for organizations seeking to operationalize a response.

MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome) [9], CSA's agentic AI threat modeling framework, provides the most directly applicable vocabulary for reasoning about AI systems operating in offensive security contexts. MAESTRO's layered approach to agentic security — examining threats at the model layer, the tool integration layer, the orchestration layer, and the environment layer — maps closely to the attack surface of AI-autonomous vulnerability discovery systems. Organizations assessing the risk of adversarial AI agents operating against their systems should use MAESTRO as a primary analytical framework. MAESTRO's treatment of prompt injection as a systemic control-flow vulnerability, rather than a simple input sanitization problem, is particularly relevant to organizations deploying AI-assisted defensive tools that interact with untrusted code and data.

The AI Controls Matrix (AICM) [10] provides a comprehensive control framework across 18 domains that addresses both the development and deployment of AI systems. For organizations deploying AI-assisted security tools — including AI-driven vulnerability scanners, automated penetration testing agents, and AI-augmented threat detection systems — the AICM's control domains for AI supply chain security, model governance, and data security provide a structured basis for assessing and improving security posture. The AICM's Shared Security Responsibility Model (SSRM) is also relevant for organizations that use third-party AI security services: understanding which controls are the responsibility of the model provider versus the deploying organization is essential when those services have access to sensitive code and system information.

CSA's Agentic AI Red Teaming Guide [11] offers operational guidance for security teams conducting adversarial testing of AI systems and using AI systems in adversarial testing. The guide's emphasis on multi-agent architectures, tool access controls, and the specific threat categories that agentic systems introduce complements the threat model in this paper and provides a practical starting point for organizations seeking to incorporate AI into their red team programs.

Zero Trust guidance from CSA addresses the architectural controls that limit blast radius when vulnerabilities are exploited — a defensive posture that becomes significantly more valuable as AI compresses exploitation timelines. The principle that no system or user should be implicitly trusted based on network location alone applies with particular force when the fastest observed adversary breakout times are measured in seconds. CSA's Zero Trust guidance provides a mature framework for implementing the segmentation and authentication controls that reduce exposure.

The STAR (Security Trust Assurance and Risk) program provides a mechanism for organizations to communicate their security posture to stakeholders, including in the context of emerging AI security capabilities. As AI-assisted vulnerability discovery becomes a standard component of security assessments, STAR-level 2 assessments that incorporate AI-specific controls provide meaningful assurance signals to customers and partners about organizational readiness for this threat environment.

Conclusions and Recommendations

The transition to AI-autonomous vulnerability discovery is not approaching — it has arrived. The research milestones of 2024 and 2025 document a world in which AI systems can discover and reproduce novel high-severity vulnerabilities at scales and speeds that no human team can match, at costs that make offensive research economically accessible to a broader range of threat actors than ever before. The governance and operational frameworks that security teams have relied on are straining to adapt, and the window for proactive investment before adversarial use of these capabilities becomes routine is narrowing.

For security leaders, the priority actions implied by this analysis are both immediate and strategic.

In the near term, organizations should compress patch deployment timelines for critical and high-severity vulnerabilities to reflect AI-accelerated exploitation windows, targeting 24 to 48 hours for internet-facing systems with known active exploitation and 7 days for high-severity findings without active exploitation evidence — timelines that exceed current CISA KEV defaults and reflect the compressed exploitation windows that AI enables. They should establish or expand SBOM programs to enable rapid identification of vulnerable dependencies when AI-discovered vulnerabilities are publicly disclosed, and review network segmentation architectures to ensure that lateral movement following initial exploitation encounters meaningful architectural barriers.

Over a 6- to 12-month horizon, organizations developing or maintaining significant software should evaluate incorporating AI-assisted security review into their SDLC, targeting the highest-risk components — memory management, authentication, and cryptographic code — with focused analysis designed for human validation. Security operations teams should develop detection engineering content specifically targeting AI-assisted attack patterns, supplementing existing rule sets with signatures derived from the characteristic

behaviors of automated exploitation pipelines. Participation in coordinated disclosure programs and structured threat intelligence sharing should be elevated as a strategic priority, particularly for organizations maintaining critical infrastructure that is likely to be targeted by AI-augmented adversaries.

At the strategic level, security leadership should engage with the governance questions this technology raises: what access controls should govern AI systems capable of autonomous vulnerability discovery, how should organizations engage with their AI security vendors on responsible disclosure policies, and how should the security workforce be developed to maintain expertise relevant to an AI-accelerated threat environment. The precedent set by Project Glasswing — restricting access to frontier offensive AI capabilities while investing in coordinated defensive use — suggests one model for managing this tension. The broader industry is still working out the norms, and organizations with meaningful standing in that conversation should participate actively.

The model is already the red team. The question now is whether defenders can adapt quickly enough.

References

[1] Anthropic Frontier Red Team. "Evaluating and mitigating the growing risk of LLM-discovered 0-days." red.anthropic.com, February 2026.

[2] Anthropic. "Project Glasswing: Securing critical software for the AI era." Anthropic, April 2026.

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

[4] Ullah et al. "From CVE Entries to Verifiable Exploits: An Automated Multi-Agent Framework for Reproducing CVEs." arXiv:2509.01835, September 2025.

[5] Deng et al. "PentestGPT: Evaluating and Harnessing Large Language Models for Automated Penetration Testing." USENIX Security Symposium 2024.

[6] CrowdStrike. "2026 CrowdStrike Global Threat Report." CrowdStrike, February 2026.

[7] Verizon. "2025 Data Breach Investigations Report." Verizon Business, 2025.

[8] Anthropic. "Coordinated Vulnerability Disclosure for Claude-Discovered Vulnerabilities." Anthropic, 2026.

[9] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 2025.

[10] Cloud Security Alliance. "AI Controls Matrix." CSA Research, 2025.

[11] Cloud Security Alliance. "Agentic AI Red Teaming Guide." CSA AI Safety Initiative, May 2025.

[12] Anthropic. "Claude Mythos Preview." red.anthropic.com, April 2026.

[13] Integsec. "In the Age of AI: The Vanishing Gap Between Vulnerability Disclosure and Exploitation." Integsec Blog, January 2026.

[14] Google. "OSS-Fuzz: Continuous Fuzzing for Open Source Software." GitHub, 2025.

