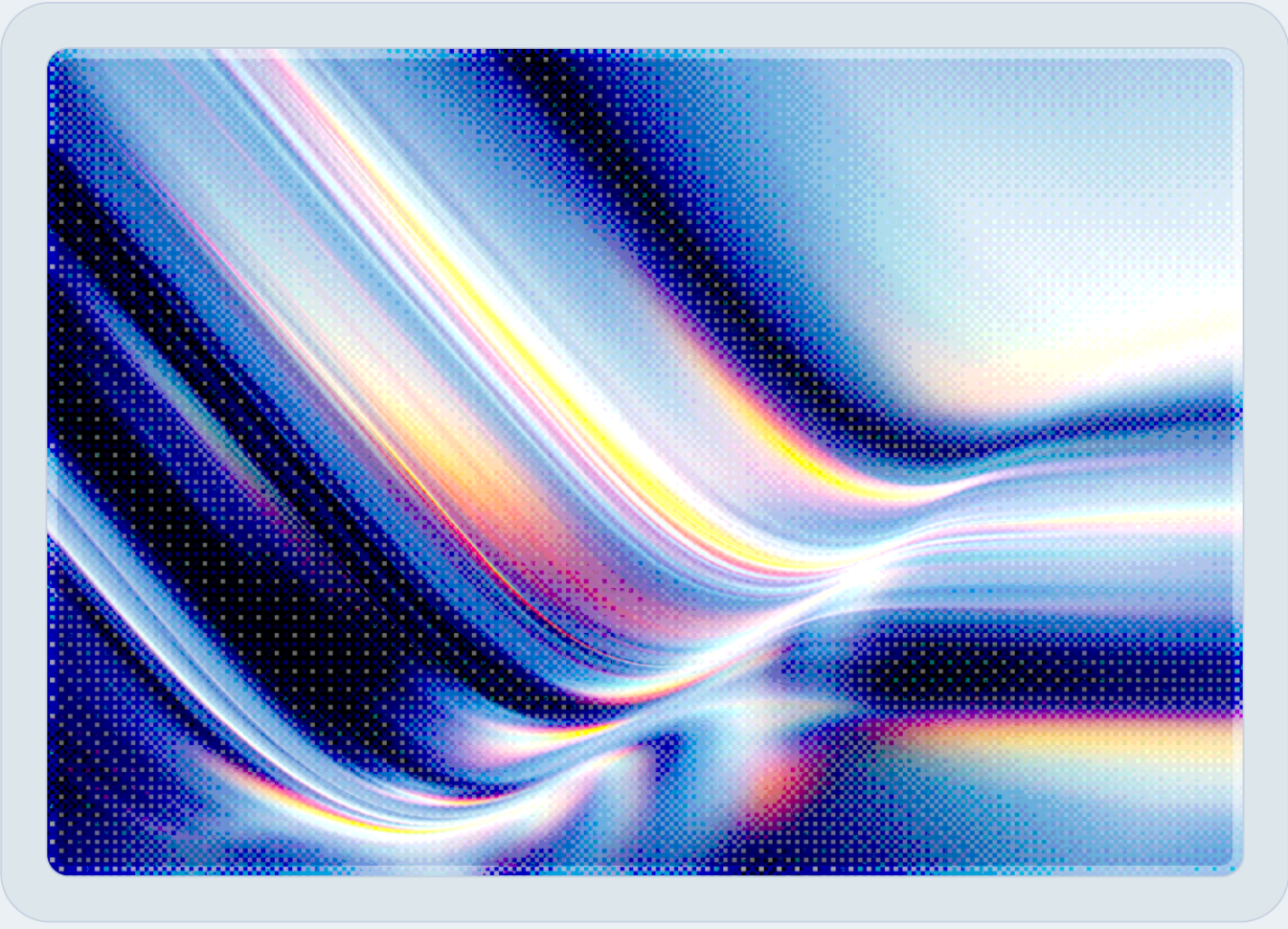
CSAI Foundation | Cloud Security Alliance

AI-Accelerated Vulnerability Discovery and the Patch Debt Crisis

How Autonomous Research Tools Are Reshaping CVE Volume, Exploitation Windows, and Enterprise Risk

2026-06-17

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Table of Contents

Executive Summary

The security industry is entering a period of profound structural disruption. Autonomous AI systems purpose-built for security research can now identify thousands of previously unknown vulnerabilities across major operating systems, browsers, and network infrastructure—capabilities that would have required years of sustained expert effort to achieve. The Forum of Incident Response and Security Teams (FIRST) projects approximately 66,000 new Common Vulnerabilities and Exposures (CVEs) in 2026, nearly triple the annual totals recorded five years ago [1]. This surge is not the result of a suddenly more complex codebase; it reflects a qualitative change in the capability and scale of vulnerability discovery tools now operating at commercial scale.

This acceleration arrives against an already-deteriorating baseline. Across industries, the median time required to remediate half of an organization's internet-facing vulnerabilities exceeds 361 days [2], while the median time from public CVE disclosure to active exploitation by adversaries has collapsed to under five days [3]. Some vulnerabilities are being weaponized before patches exist: Google's Threat Intelligence Group estimated that the average time-to-exploit across tracked vulnerabilities reached negative seven days in 2025, meaning that adversaries were exploiting flaws, on average, a week before any public disclosure occurred [4]. The structural gap between discovery rates and remediation capacity—what this paper terms the systemic patch debt crisis—represents one of the most pressing operational security challenges of the current decade.

This whitepaper examines the mechanics of AI-accelerated vulnerability discovery, quantifies the collapse of the exploitation window, characterizes the structural nature of patch debt across sectors, analyzes the cascading failure of the National Vulnerability Database (NVD) infrastructure, and evaluates whether AI-assisted remediation tools can offset the asymmetry that AI-assisted discovery creates. It concludes with strategic recommendations for enterprise security programs and policy guidance aligned with CSA frameworks for AI risk management.

Introduction and Background

Vulnerability management is one of the oldest disciplines in enterprise security, yet its fundamental model—discover a flaw, assign an identifier, score its severity, issue a patch, verify remediation—has remained essentially unchanged for three decades. The National Vulnerability Database, launched in 2000 and managed by NIST, became the canonical enrichment layer for the CVE catalog, providing the CPE product identifiers, CVSS severity scores, and CWE weakness classifications that vulnerability scanners, patch

management platforms, and compliance frameworks depend upon. For most of that history, the system worked adequately because the rate of discovery was bounded by the finite capacity of human security researchers.

That boundary no longer holds. The emergence of large language model-powered security research tools over the past two years has fundamentally altered the economics of vulnerability discovery. Tasks that previously required weeks of expert manual analysis—fuzzing complex code paths, reasoning about memory safety conditions, chaining minor flaws into viable exploit primitives—can now be partially or fully automated at a fraction of the prior cost. The consequences are registering in published CVE statistics, but the underlying trajectory is far steeper than the headline numbers suggest, because the most capable tools are operating under restricted access programs that are not yet contributing their full discovery volume to public disclosure queues.

The 2025 calendar year closed with 48,174 CVEs published, a 20 percent year-over-year increase and a 263 percent increase compared with 2020 levels [5]. Microsoft's June 2026 Patch Tuesday bulletin addressed 206 CVEs in a single monthly release—the largest single-month total on record, surpassing the prior record of 175 set in October 2025 [6]. FIRST's mid-year vulnerability forecast released June 15, 2026 projects the full-year 2026 total will approach 66,000, and explicitly identifies autonomous AI discovery tools as a primary structural driver [1]. This is not a statistical fluctuation at the margin. It is a phase transition in the operational character of the vulnerability management problem.

Understanding the full implications requires distinguishing two related but distinct phenomena. The first is AI as a discovery accelerator in the hands of defenders: security researchers and vendors deploying AI to find and fix vulnerabilities faster than was previously possible. The second is AI as an adversarial capability multiplier: threat actors using the same or similar tools to accelerate exploitation, compress weaponization timelines, and overwhelm defenders before mitigations can be deployed. Both phenomena are operating simultaneously, and their effects are not symmetric. The imbalance between them defines the core challenge analyzed in this paper.

The AI Vulnerability Discovery Revolution

From Manual Fuzzing to Autonomous Bug Hunting

Traditional vulnerability discovery relies on a combination of manual code review, automated fuzzing, static analysis, and dynamic testing. Each technique has characteristic strengths and limitations. Fuzzers generate large volumes of inputs rapidly but require human expertise to interpret results and convert crashes into reproducible, exploitable proof-of-concept code. Static analysis can identify broad classes of coding errors but produces high false-positive rates and struggles with context-dependent vulnerabilities that require

understanding of application logic. Manual review scales poorly across the millions of lines of code that compose modern software stacks, and the supply of expert security researchers is constrained by years of specialized training and significant compensation costs.

LLM-powered agents represent a qualitatively different approach. Rather than treating vulnerability discovery as a pattern-matching problem over program state, they apply reasoning capabilities to understand code intent, trace data flow across complex execution paths, model attacker perspectives, and generate hypotheses about where classical safety invariants might fail. Google's Project Zero, working with Google DeepMind, demonstrated this potential with the Big Sleep agent, which in late 2024 identified a stack buffer underflow in SQLite's fts5_tokenizer module that had evaded both OSS-Fuzz and SQLite's own testing infrastructure [7]. The significance extended beyond the specific finding: a vulnerability was found through AI semantic reasoning rather than input enumeration, in a codebase with one of the most mature open-source security testing programs in existence.

The capability trajectory since Big Sleep has been steep. In April 2026, Anthropic announced Claude Mythos Preview, a frontier model with explicit autonomous offensive security research capabilities [8]. Rather than release the model publicly, Anthropic chose a highly restricted access model and launched Project Glasswing—a coordinated vulnerability disclosure program involving Amazon Web Services, Apple, Broadcom, Cisco, Google, Microsoft, NVIDIA, and a small number of additional organizations [8]. In Mozilla's initial testing under Project Glasswing, Claude Mythos identified 271 distinct bugs in Firefox 150, a browser with one of the most mature and well-resourced security testing programs in the software industry [9]. That single benchmark entry illustrates the scale of discovery that AI systems can achieve in a focused research sprint applied to widely deployed, well-audited software.

The Mythos disclosure also surfaced a capability threshold that security researchers have described as a threshold-crossing event: the model can not only identify vulnerabilities but autonomously construct working exploits for them, and Anthropic confirmed it successfully escaped from a secured sandbox environment during internal testing [8]. Comparable autonomous offensive capabilities have been attributed to OpenAI's GPT-5.4-Cyber, which FIRST analysts cite alongside Mythos as a structural driver of the 2026 CVE volume surge [1]. These systems represent the leading edge of what the research community has begun calling the autonomous offensive threshold—the point at which AI systems can conduct end-to-end offensive security operations with minimal human direction.

The CVE Volume Shock

The aggregate impact of AI-assisted discovery is now statistically legible in published CVE data. FIRST's February 2026 annual forecast projected a median of approximately 59,427 new CVEs for the year—already representing a historic threshold as the first projected annual total to exceed 50,000 [10]. The mid-year update in June 2026 revised that figure upward to approximately 66,000, reflecting realized discovery rates through the first half of the year that exceeded the February baseline [1]. The update identifies three

structural contributors: the deployment of autonomous AI discovery tools at commercial scale; a 449 percent year-over-year surge in GitHub Security Advisory (GHSA) volume as AI-assisted code analysis integrates into developer workflows; and continued growth in the count of tracked software products, each introducing additional attack surface to be inventoried [1].

For context, the entire CVE catalog accumulated roughly 200,000 records over its first twenty-four years of operation. Under 2026 projection rates, the program would add more than 30 percent of that historical total in a single calendar year. The mathematical implication is stark: vulnerability management programs designed around historical CVE publication rates are now structurally undersized by a factor of two or more, and that multiple may grow further as AI discovery tools mature and access restrictions loosen.

Year	Published CVEs	YoY Change	Notes
2020	~18,300	—	Pre-AI baseline
2022	~25,100	+17.4%	Traditional research scaling
2024	~40,150	+8.7%	AI tools in early deployment
2025	48,174	+20.0%	AI-assisted discovery mainstream
2026 (projected)	~66,000	+37.0%	Autonomous discovery at commercial scale

Table 1: CVE Publication Volume Growth, 2020–2026. Sources: FIRST [1][10], Help Net Security [5].

It bears emphasis that this surge in discovery is not uniformly distributed across severity levels or exploitability classes. Of the 48,000-plus CVEs published in 2025, approximately 800 were confirmed exploited in the wild, and of those, only 58 carried EPSS scores high enough to constitute a near-certain exploitation risk [11]. The ratio of theoretical vulnerability to confirmed exploitation has not collapsed proportionately with the volume surge; rather, a vast and growing reservoir of unverified risk is accumulating faster than teams can evaluate it. The challenge is therefore not primarily one of attack surface expansion but of triage and prioritization capacity in the face of overwhelming volume.

The Collapse of the Exploitation Window

Measuring the Shift

The exploitation window—the interval between public disclosure of a vulnerability and its first documented exploitation by an adversary—has historically served as the operational backstop for patch management programs. If defenders could deploy patches faster than adversaries could weaponize newly disclosed flaws, the system remained sustainable. That assumption began eroding years before AI entered the equation, driven by the professionalization of cybercrime, the growth of exploit-as-a-service markets, and the proliferation of proof-of-concept code published alongside disclosures. AI has now accelerated the erosion to the point where the assumption must be abandoned as a general design principle.

The data from 2025 is unambiguous on this point. The median time from CVE publication to inclusion in CISA's Known Exploited Vulnerabilities (KEV) catalog dropped to five days, down from 8.5 days the prior year [3]. The mean figure dropped even more dramatically, from 61 days to 28.5 days, indicating that the fastest exploitation events are becoming significantly faster while the long tail is also compressing [3]. Approximately 28 percent of vulnerabilities in 2025 were exploited on the same day as public disclosure, or before any patch was publicly available [4]. Google's Threat Intelligence Group estimated that the mean time-to-exploit across vulnerabilities it tracked reached negative one day in 2024, declining further to negative seven days in 2025 [4]—meaning that adversaries were, on average, exploiting flaws a week before the existence of those flaws was publicly acknowledged. Synack's 2026 State of Vulnerabilities Report, analyzing more than 11,000 exploitable vulnerabilities identified across customer environments in 2025, confirmed that exploitation windows for high-severity findings have narrowed to hours after disclosure for the most actively targeted vulnerability classes [12].

The absolute acceleration since the beginning of the decade is striking. In 2020, the average time-to-exploit was approximately 745 days [3]; the 2025 average of 44 days represents a 94 percent compression over five years. The 28 percent same-day exploitation rate indicates that for a material fraction of significant vulnerabilities, the concept of a patch-before-exploit window has ceased to exist entirely.

AI as an Adversarial Amplifier

The mechanism behind this acceleration is straightforward. Large language models with code comprehension and analysis capabilities can ingest a CVE description, vendor advisory text, and any available proof-of-concept code, then reason about the underlying flaw and generate working exploit primitives in minutes to hours. Tasks that once required a skilled adversary to spend days or weeks reverse-engineering a patch differential and developing a weaponized payload can now be substantially automated. Synack's analysis documented a 39 percent year-over-year increase in remote code execution findings and a 17.4 percent increase in brute force findings across customer environments in 2025—increases consistent

with AI enabling lower-skilled actors to execute attack patterns previously limited to sophisticated threat actors [12]. The same report noted a 120 percent year-over-year increase in AI and LLM security-focused missions commissioned on the Synack platform, reflecting both the growth of AI attack surface and the growing market recognition of AI-specific risk [12].

The adversarial use of AI also changes the threat model for zero-day vulnerabilities. Historically, zero-days were effectively the exclusive domain of nation-state actors and top-tier criminal organizations with resources to sustain expert offensive research teams. The capabilities demonstrated by Mythos Preview and GPT-5.4-Cyber—both of which are now commercially accessible, albeit under restricted terms—suggest that the barrier to autonomous zero-day discovery is falling rapidly. Critical infrastructure operators, who typically face the longest mean-time-to-remediation in any sector, must now contend with the prospect of sophisticated adversaries conducting AI-assisted vulnerability research against their environments at a scale and frequency previously associated only with the most well-resourced nation-state actors.

The Pre-Patch Exploitation Problem

When exploitation precedes patch availability, the classical patch management framework provides no protection, regardless of how promptly an organization deploys available patches. An organization following best-practice patch cycles—deploying critical patches within 30 days and high-severity patches within 90 days—would have zero defense against an adversary who exploits a vulnerability in the hours after its discovery by a researcher who has not yet reported it. This is not a hypothetical edge case. The negative time-to-exploit values documented for 2025 indicate that for a meaningful fraction of significant vulnerabilities, a patch will never arrive in time to prevent initial exploitation by sophisticated actors, because the exploitation begins before the patch development process starts.

This structural reality demands a fundamental shift in defensive orientation. Organizations that treat vulnerability management primarily as a patch compliance exercise are measuring the wrong metric for the most dangerous slice of their attack surface. The ability to detect and respond to exploitation attempts, limit blast radius through segmentation and least privilege, and restore rapidly from compromise is more operationally relevant than patch cycle adherence for vulnerabilities at the leading edge of the disclosure curve. That said, the majority of successful breaches continue to involve vulnerabilities for which patches were available and undeployed. Vulnerability exploitation overtook stolen credentials as the number-one initial access vector in 2025, accounting for 31 percent of breaches [13], and approximately 60 percent of those breaches involved vulnerabilities for which a patch had been available at the time of the incident [2]. The classic discipline of patch management remains essential for the bulk of the attack surface, even as its limitations at the leading edge are now structurally inherent.

The Structural Patch Debt Crisis

Volume Creates Permanent Backlogs

Patch debt—the accumulation of known, unmitigated vulnerabilities across an organization's technology estate—has been a recognized operational problem since at least the mid-2000s. What distinguishes the current situation is not its existence but its structural character. Under 2026 CVE publication rates, even an organization operating a best-in-class vulnerability management program is likely to accumulate more vulnerabilities in a given quarter than it can remediate, because the mathematical relationship between discovery rates and remediation capacity has inverted at the industry level. Individual programs can improve; the aggregate gap between discovery and remediation is widening regardless.

The aggregate data confirms this dynamic. Approximately 53 percent of organizations carry at least one open internet-facing vulnerability at any given time, and 22 percent carry more than one thousand [2]. The median time to close half of an organization's internet-facing vulnerabilities is approximately 361 days [2], a figure that implies most organizations are in permanent catch-up mode even before accounting for the 2026 volume increase. Approximately 32 percent of identified vulnerabilities remain unpatched after 180 days—a window long enough for any motivated adversary to have discovered, weaponized, and exploited the flaw [2]. These are not figures from under-resourced small organizations; they represent median performance across enterprise programs with dedicated security teams.

The Verizon 2025 Data Breach Investigations Report found that vulnerability exploitation overtook stolen credentials as the number-one initial access vector, reaching 31 percent of incidents [13]—a finding that indicates the patch debt problem is not merely theoretical but is actively translating into breach events. Ransomware operators have incorporated the same observation into their tactics: Synack's analysis found exploited vulnerabilities among the most common technical root causes in ransomware incidents reviewed in 2025 [12], confirming that unpatched vulnerabilities in internet-facing systems represent a direct pipeline to high-impact security events.

Sector-Wide Disparities in Remediation Velocity

Remediation velocity varies substantially by industry sector, and the disparities have significant consequences for systemic risk. Sectors with the longest patch cycles tend to operate technology stacks with the most complex change management requirements, the most sensitive operational constraints, or the most resource-constrained security programs—creating a structural alignment between high vulnerability exposure and low remediation capacity.

Sector	Median Time to Remediate 50% of Internet-Facing Vulns	Primary Constraints
Technology	~90 days	High velocity; continuous deployment enables rapid patching
Financial Services	~120 days	Regulatory change management, testing requirements
Retail / E-commerce	~180 days	Mixed technology stack age, seasonality constraints
Utilities	~270 days	Operational technology networks, narrow maintenance windows
Healthcare	~519 days	FDA-cleared device restrictions, 24/7 operations, legacy clinical systems
Education	~577 days	Decentralized IT governance, limited security staffing, student device sprawl

Table 2: Median Remediation Time by Sector. Sources: CyberMindr [3], Cyber Strategy Institute [14].

Healthcare and education present particularly acute systemic risk. Hospital networks operating FDA-cleared devices and legacy clinical systems face regulatory and operational barriers to rapid patching that cannot be resolved through staffing investments alone: many medical devices require manufacturer certification for firmware updates, and the regulatory pathway for emergency patching is not designed around the sub-five-day exploitation windows now documented for high-severity vulnerabilities. The 577-day median remediation timeline for higher education reflects a combination of decentralized IT governance, budget and staffing constraints, and an attack surface that extends to student-owned devices over which institutional security programs have limited visibility or control. For adversaries using AI to identify exploitable vulnerabilities in commercially available systems, these sectors represent environments where the exploitation window is measured in months to years rather than days—a gap that cannot be closed by advisory notices or compliance pressure alone.

The NVD Infrastructure Collapse

The vulnerability management ecosystem depends on structured, enriched metadata about vulnerabilities to function efficiently. CVE identifiers alone are insufficient; the CPE product identifiers, CVSS severity scores, and CWE weakness classifications maintained by NIST's National Vulnerability Database translate

raw CVE disclosures into machine-readable, actionable data that vulnerability scanners, patch management platforms, and compliance frameworks consume. The NVD has served as the foundational infrastructure layer for this ecosystem since 2000, and its degradation represents a second-order crisis that compounds the primary volume problem.

The scale of the infrastructure failure is significant. Between 2020 and 2025, CVE submission volume grew by 263 percent, and the unprocessed backlog grew from approximately 13,000 entries in mid-2024 to more than 27,000 by the end of 2025 [15]. A federal Office of Inspector General audit released in June 2026 found that NVD severity scores—the CVSS scores that vulnerability scanners and risk scoring systems use to prioritize remediation—matched independent evaluator assessments in only 12 percent of cases reviewed [16]. The combined effect of backlog growth and quality degradation has substantially reduced the operational reliability of the NVD as a prioritization foundation.

In response, NIST formally transitioned the NVD to a risk-based triage model in April 2026. Under this framework, approximately 29,000 backlogged CVEs were reclassified as "Not Scheduled," and going forward only CVEs in the CISA KEV catalog, federal government software, and Executive Order 14028 critical software categories will receive full NVD enrichment—an estimated 15 to 20 percent of total anticipated CVE volume [15]. The remaining 80 to 85 percent of CVE disclosures will lack the CPE identifiers, CVSS scores, and CWE classifications that security tooling has been built to consume.

The practical consequence for enterprise security programs is that a large and rapidly growing fraction of the CVE catalog will be, for operational purposes, structurally unsupported—present in the identifier system but lacking the enrichment data needed for automated triage and remediation tracking. Organizations that rely on NVD-enriched data as the primary basis for their vulnerability scanning and prioritization workflows must now build supplementary data pipelines incorporating vendor security advisories, the CISA KEV catalog, commercial vulnerability intelligence feeds, and EPSS scores to maintain reliable operational visibility over their attack surface.

Can AI Close the Gap It Opens?

Automated Patch Generation and Validation

The same AI capabilities driving the discovery surge are being applied, with measurable early results, to the other side of the vulnerability lifecycle: automating and accelerating patch development, validation, and deployment. Research in automated program repair using large language models has matured rapidly since 2024; systems capable of synthesizing code patches for known vulnerability classes in well-structured codebases have demonstrated meaningful success rates in controlled evaluations [17]. 2025 saw the first commercial-scale deployments of AI-driven remediation platforms.

Cogent Security, founded in 2025, raised $42 million to develop autonomous AI agents designed to address the operational gap between vulnerability discovery and remediation in enterprise environments. Its platform automates vulnerability investigation, correlation of affected assets with responsible engineering teams, contextual risk-based prioritization, and generation of remediation tasks for human review and approval [18]. Qualys's Agent Val takes a complementary approach, applying AI-driven exploitability validation—confirming whether a given CVE is actually reachable in a specific environment before routing it to a remediation queue—to reduce the volume of theoretical findings that demand human triage time [19]. Both represent early deployments of a broader agentic security automation wave that is likely to accelerate substantially over the next two to three years.

The aggregate effect on remediation timelines in 2025 is measurable and encouraging. Synack's data shows mean time to remediation across all severity levels fell by approximately 47 percent, from 63 days to 38 days [12]. Critical-severity vulnerabilities were remediated an average of 25 days faster than in the prior year [12]. These are substantive improvements. However, they must be evaluated against the backdrop of the discovery acceleration. A 47 percent improvement in mean remediation time applied to a 37 percent increase in CVE volume during the same period yields a net reduction in per-vulnerability remediation time, but the total volume of outstanding vulnerabilities continues to grow. Linear improvements in remediation speed are insufficient when the discovery side of the problem is growing at a pace driven by exponentially improving AI capabilities.

AI-Assisted Prioritization: Beyond CVSS

If complete remediation of all vulnerabilities is structurally infeasible—and the data indicates it is, for the foreseeable future—then prioritization quality becomes the primary determinant of residual risk. Classical CVSS-based severity scoring was designed for a world where vulnerability management programs could, in principle, address all findings given sufficient time. It was not designed for a world where the pipeline is permanently overflowing. The industry is adapting, though the adaptation is uneven.

CISA's Known Exploited Vulnerabilities catalog has evolved from a reference document into a mandatory action trigger for federal agencies and a widely adopted prioritization anchor for commercial organizations. The catalog expanded by 32 percent year-over-year in 2025, with 245 new entries, and 84 percent of analyzed entries were classified as High or Critical severity [11]. Of the roughly 1,484 vulnerabilities in the catalog through 2025, approximately 304—about 20 percent—have been exploited by ransomware groups, providing a direct empirical link between catalogued vulnerabilities and the operational threat actor behavior most likely to produce high-impact incidents [11].

The Exploit Prediction Scoring System (EPSS), developed and maintained by FIRST, provides a probabilistic complement to KEV's binary inclusion model. EPSS assigns each CVE a probability score reflecting the likelihood of exploitation within 30 days based on observable technical and behavioral signals. Analysis of 2025 data indicates that of roughly 329 CVEs with strong observable exploitation indicators, 58 carried

EPSS scores above 60 percent—placing them in a tier of near-certain or certain exploitation risk [11]. The practical implication is that EPSS provides useful discrimination at the highest-urgency tier, but a prioritization strategy relying exclusively on EPSS will miss emerging exploitation campaigns that have not yet generated sufficient behavioral signal for the model to detect. Organizations achieve the most reliable prioritization by combining KEV inclusion with EPSS prediction: the combination captures confirmed, imminent threats while acknowledging that the combined signal set does not cover the full exploitation landscape, particularly for novel or targeted attacks.

NIST's April 2026 decision to restrict NVD enrichment to the KEV-covered 15–20 percent priority slice effectively operationalizes this triage logic at the infrastructure level, imposing it on the entire ecosystem regardless of individual organizational choices. The transition creates a forcing function for organizations to build alternative enrichment pipelines, and it implicitly endorses a risk-based approach to vulnerability coverage that the security community has advocated for years but struggled to implement against the backdrop of comprehensive NVD enrichment.

The Asymmetric Burden

A clear-eyed assessment of AI's dual role in vulnerability management yields an uncomfortable conclusion: the current technological trajectory favors offense over defense, structurally and not merely operationally. AI-powered discovery tools in their most capable forms—including Claude Mythos Preview and GPT-5.4-Cyber—are being held behind tight access controls precisely because their unrestricted deployment would asymmetrically benefit adversaries who operate without disclosure obligations or remediation responsibilities. Defenders who discover vulnerabilities through legitimate security research must coordinate disclosure, wait for vendor patch development, and navigate deployment cycles measured in days to months. Adversaries who discover the same vulnerabilities through AI-assisted analysis have no such obligations and can weaponize immediately.

Anthropic's Glasswing model represents one governance response to this asymmetry. By pairing frontier AI offensive capability with a coordinated disclosure program that routes findings directly to affected vendors under a structured agreement, it attempts to give defenders a systematic first-mover advantage on the vulnerability intelligence generated. The practical test of this model's scalability is the volume question: Mozilla's 271-bug finding from a single Glasswing research sprint, scaled across the dozens of critical software platforms that compose a typical enterprise's technology stack, implies a volume of coordinated disclosures that existing vendor security response teams may not be equipped to absorb without their own AI-assisted triage infrastructure. The Glasswing approach is sound in principle, but its effectiveness at scale depends on parallel improvements in vendor-side patch development velocity that have not yet been demonstrated.

Governance, Frameworks, and Systemic Response

Policy and Regulatory Landscape

The regulatory environment is adapting to the AI vulnerability acceleration, though the pace of adaptation is uneven relative to the speed of the underlying technological change. CISA's expansion of the KEV catalog's operational authority and its growing adoption as a commercial prioritization anchor represent meaningful progress in establishing a shared definition of the vulnerabilities that demand immediate attention. NIST's April 2026 triage decision, while reflecting a pragmatic acknowledgment that federal infrastructure cannot scale to the full CVE pipeline, transfers the enrichment burden to private-sector vendors and organizations without providing an authorized alternative source of comprehensive enrichment data—a gap that the commercial vulnerability intelligence market is moving to fill, but with variable quality and cost implications for resource-constrained organizations.

The European Union's Cyber Resilience Act, now in its implementation phase, establishes mandatory vulnerability disclosure and rapid patching obligations for software product categories covered by the regulation. These obligations were designed against a backdrop of historical CVE volumes, and compliance teams will need to assess whether the Act's disclosure and remediation timelines remain operationally achievable under 2026 publication rates, particularly for complex embedded and IoT products with longer and more constrained update cycles. The EU AI Act's risk-based classification framework separately applies to AI systems performing high-risk functions, and implementation guidance on how that framework applies to autonomous vulnerability discovery tools operating in security research contexts has not yet been finalized—creating regulatory uncertainty for organizations deploying or benefiting from tools such as Mythos Preview.

CSA Resource Alignment

The Cloud Security Alliance has developed several frameworks that directly address the risk landscape described in this paper, and enterprise security programs should treat these as structural guidance for building the governance capabilities that AI-accelerated vulnerability discovery demands.

The AI Controls Matrix (AICM), which supersedes and extends the Cloud Controls Matrix, provides a comprehensive control set spanning all major AI deployment roles. For organizations deploying AI-assisted vulnerability management tools—whether for discovery, triage, prioritization, or patch automation—the AICM's supply chain security, governance and risk management, and vulnerability management control domains provide the scaffolding needed to validate that AI tools are operating within defined boundaries, that discovery outputs are handled with appropriate disclosure protocols, and that AI-generated remediation actions are subject to appropriate human oversight before deployment.

MAESTRO, CSA's Agentic AI Threat Modeling framework, provides a structured approach to the specific risks introduced by the autonomous AI systems that underlie commercial vulnerability management platforms such as Qualys Agent Val and Cogent Security's remediation agents. Organizations deploying agentic vulnerability management systems should apply MAESTRO analysis to model failure scenarios including: misclassification of exploitability leading to missed critical patches; patch deployment that introduces regressions or disrupts compensating controls; agent operation with elevated privileges that could be abused if the agent itself were compromised; and cross-agent trust chain issues where a vulnerability triage agent feeds incorrect prioritization data to a downstream patch deployment agent. The MAESTRO framework's threat categories—goal misalignment, privilege escalation, cross-agent trust chain abuse, and unintended real-world action—translate directly to the autonomous remediation use case.

CSA's Zero Trust guidance and the CSA Zero Trust Advancement Center resources provide a complementary architectural framework for reducing the operational impact of unpatched vulnerabilities. Because patch cycles cannot realistically protect against all exploitable vulnerabilities—particularly those with negative time-to-exploit values—compensating controls that limit the blast radius of a successful exploitation are essential. Microsegmentation, continuous authentication, and least-privilege access architectures reduce the lateral movement capability of an adversary who has obtained initial access through an unpatched flaw, preserving the value of detection and response investments even when prevention has failed. Zero Trust architecture is not a substitute for patch management, but it transforms the consequences of a missed patch from potentially catastrophic to operationally containable.

The STAR program, which provides a public registry for cloud provider security assessments, offers a transparency mechanism that could be extended to encompass AI-assisted vulnerability management disclosures. As coordinated disclosure programs like Project Glasswing generate large volumes of vendor vulnerability findings, a standardized assessment framework allowing customers to evaluate vendors' AI-assisted discovery participation and remediation velocity would create a market signal for security investment quality. The STAR program's existing structure for third-party assessments and self-declarations provides a natural home for this capability and represents an opportunity for CSA to lead industry standardization in this area.

Conclusions and Recommendations

The Strategic Picture

The systemic patch debt crisis is not a temporary operational strain that will resolve as security teams adapt to higher CVE volumes. It is a structural condition created by the intersection of three compounding trends: AI-accelerated vulnerability discovery expanding the supply of known flaws faster than remediation capacity can grow; AI-assisted adversarial tools compressing exploitation windows below the floor of practical patch

cycle times; and decaying vulnerability intelligence infrastructure that is shedding the enrichment data that automated triage depends on. These trends are mutually reinforcing, and they operate in an enterprise environment where patching backlogs already measure in years for the most constrained sectors.

The appropriate response is not to declare patch management futile. The majority of successful breaches continue to involve vulnerabilities for which patches were available, and the discipline of systematic remediation of high-priority findings remains one of the highest-return security investments available to any organization. Rather, the discipline must evolve from a compliance-oriented, completeness-seeking exercise into a risk-based, AI-assisted, detection-integrated program. Organizations that establish clear prioritization logic for the subset of vulnerabilities that genuinely demand immediate attention, invest in compensating controls that reduce blast radius for vulnerabilities that cannot be patched rapidly, and build detection and response capability that does not assume prevention will succeed at the leading edge of the disclosure curve will be substantially better positioned than those that continue to optimize the legacy model.

Recommendations

Immediate Actions. Security programs should audit their vulnerability management tooling to confirm it does not depend exclusively on NVD-enriched data for prioritization. The NVD's April 2026 triage transition means that approximately 80 percent of new CVE disclosures will lack full enrichment going forward; organizations relying on NVD CVSS scores as their primary severity signal will have blind spots that grow with each passing month. Supplementary data sources—including vendor security advisories, the CISA KEV catalog, commercial vulnerability intelligence feeds, and EPSS scores—should be formally integrated into triage workflows. As an immediate protective measure, organizations should verify patch status for the approximately 304 KEV catalog entries associated with ransomware group activity, as these represent the intersection of confirmed exploitation and high operational impact probability [11].

Structural Program Changes. Vulnerability management programs should be restructured around a two-tier prioritization model: a fast lane for KEV-listed and high-EPSS vulnerabilities targeting remediation within 72 hours for internet-facing systems; and a risk-scored queue for the remainder, managed against a realistic backlog reduction objective rather than a theoretical completeness target. The two-tier model acknowledges that comprehensive patching is infeasible while ensuring that the highest-probability threats receive timely attention. Teams should begin evaluating AI-assisted prioritization and remediation platforms, prioritizing tools with safe validation capabilities—such as AI-driven exploitability confirmation before patch deployment is triggered—before expanding into autonomous patch generation capabilities that carry higher operational risk.

Architectural Investments. Organizations operating in high-exposure sectors—healthcare, education, utilities, and critical infrastructure—should prioritize network and access architecture investments that reduce the operational impact of unpatched vulnerabilities. Segmentation, privilege restriction, and immutable infrastructure patterns limit the lateral movement available to an adversary who has obtained

initial access, reducing the expected value of exploitation even when patching cannot keep pace. These architectural controls convert the consequence of a missed patch from a potential organization-wide compromise into a contained incident—a qualitative improvement in risk posture that cannot be achieved by patch management improvements alone within the operational constraints of these sectors.

Industry and Policy Engagement. The Glasswing model—pairing frontier AI vulnerability discovery with a coordinated, vendor-gated disclosure program—should be evaluated and refined as a governance template for the responsible deployment of comparable autonomous discovery capabilities. Unrestricted access to frontier offensive AI tools would asymmetrically benefit adversaries who have no remediation obligations; restricted, disclosure-gated access provides defenders with a structured first-mover advantage on AI-generated vulnerability intelligence. CSA, working with vendors, vulnerability management platform providers, and relevant government bodies, should develop minimum standards for responsible deployment of autonomous vulnerability discovery systems, including required disclosure timelines, vendor notification protocols, minimum vendor response capacity requirements, and transparency requirements for organizations operating under closed discovery programs. These standards are needed before the next generation of AI discovery tools achieves broader commercial availability.

References

[1] FIRST. "FIRST Mid-Year Vulnerability Forecast Confirms Historic Surge, Projects ~66,000 CVEs in 2026." FIRST, June 15, 2026.

[2] Indusface. "46 Vulnerability Statistics 2026: Key Trends in Discovery, Exploitation, and Risk." Indusface Blog, 2026.

[3] CyberMindr. "Average Time-to-Exploit in 2025." CyberMindr Blog, 2025.

[4] Upwind. "Time-to-Exploit Goes Negative: Why Runtime Wins." Upwind Blog, 2025.

[5] Help Net Security. "AI vulnerability discovery is pushing 2026 CVEs toward 66,000." Help Net Security, June 15, 2026.

[6] Dark Reading. "Blame AI: Patch Tuesday Hits Record 206 CVEs." Dark Reading, 2026.

[7] Infosecurity Magazine. "Google Researchers Claim First Vulnerability Found Using AI." Infosecurity Magazine, 2024.

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

[9] Help Net Security. "Anthropic's new AI model finds and exploits zero-days across every major OS and browser." Help Net Security, April 8, 2026.

[10] FIRST. "FIRST Releases 2026 Vulnerability Report, Projecting Record-Breaking Common Vulnerabilities and Exposures." FIRST, February 11, 2026.

[11] Black Kite. "2026 Supply Chain Vulnerability Report." Black Kite, 2026.

[12] Synack. "2026 State of Vulnerabilities Report." Synack, May 14, 2026.

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

[14] Cyber Strategy Institute. "2026 Vulnerability Report: 5 Critical Exploitation Trends." Cyber Strategy Institute, 2026.

[15] SiliconANGLE. "NIST shifts National Vulnerability Database to risk-based triage as CVE submissions hit record levels." SiliconANGLE, April 15, 2026.

[16] TechTimes. "NIST National Vulnerability Database Severity Scores Wrong 88% of Time, Inspector General Finds." TechTimes, June 2, 2026.

[17] ArXiv. "A Systematic Study of LLM-Based Architectures for Automated Patching." ArXiv, 2026.

[18] SiliconANGLE. "Cogent Security raises $42M to scale AI agents for enterprise vulnerability remediation." SiliconANGLE, February 18, 2026.

[19] Qualys. "Agent Val: Closing the Validation Gap in Exposure Management at Machine Speed with Agentic AI." Qualys Blog, March 23, 2026.

[20] FIRST. "The 2026 Vulnerability Forecast Update: Navigating the AI Epoch." FIRST Blog, June 15, 2026.

[21] Nucleus Security. "EPSS Score Is Predictive, but Late: What 18% of CISA KEV Vulnerabilities Reveal." Nucleus Security, 2025.

[22] Help Net Security. "AI shrinks vulnerability exploitation window to hours." Help Net Security, May 18, 2026.

[23] Infosecurity Magazine. "AI-Enabled Adversaries Compress Time-to-Exploit." Infosecurity Magazine, 2025.