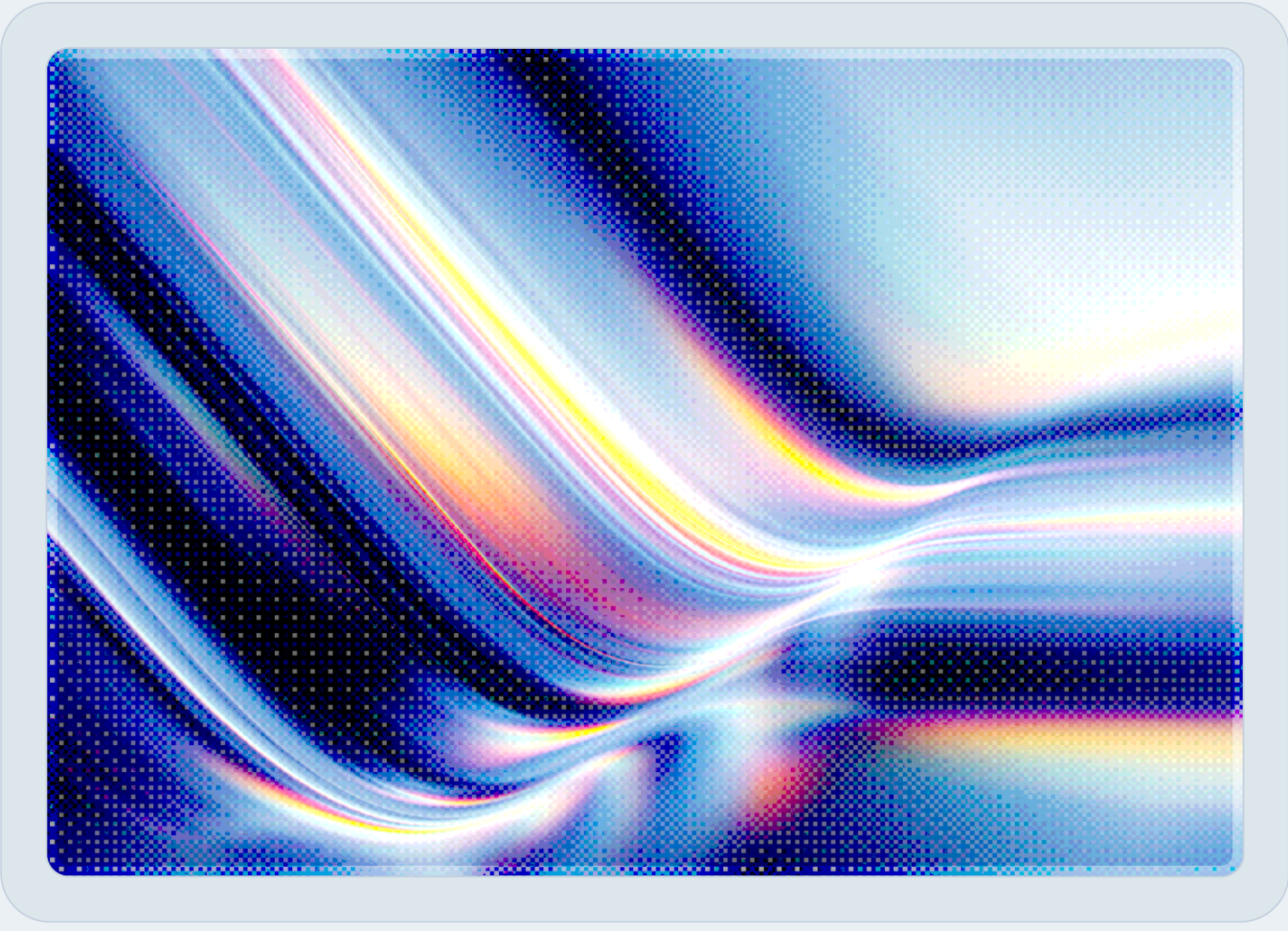
CSAI Foundation | Cloud Security Alliance

The Exploit-Before-Patch Gap

How AI-Accelerated Vulnerability Discovery Is Inverting the Defense Timeline

2026-05-14

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Table of Contents

Executive Summary

The enterprise security model that has governed vulnerability management for two decades rests on a foundational assumption: public disclosure of a vulnerability precedes exploitation at scale, and organizations have a meaningful window of time to deploy patches before attackers develop and deploy working exploits. This assumption is no longer reliable.

Three converging trends are dismantling the patch-window model. First, the interval between vulnerability disclosure and first exploitation has collapsed from measured in months to measured in days, and in many cases in hours. Mandiant's longitudinal tracking shows the median time-to-exploit fell from sixty-three days in 2018–2019 to approximately five to seven days by 2023, a trajectory that subsequent data indicates has continued downward [1][2]. Second, AI-augmented exploit development has dramatically reduced the skill and labor required to convert a disclosed vulnerability into a working weapon, democratizing a capability that was previously restricted to nation-state actors and sophisticated criminal groups. Third, and most consequentially, a growing fraction of vulnerabilities are being exploited before CVE numbers are even assigned or patch advisories published — creating what analysts have termed a "negative time-to-exploit" scenario in which the classic disclosure-patch-defend sequence never gets started.

These trends do not simply shorten the patch window. They invert the defensive model itself. When exploitation precedes disclosure, patching cannot be the primary defensive posture, because there is nothing to patch against. When AI systems can generate functional exploits from published advisories in minutes and operationalize them at machine scale, the traditional assumption that attackers require human expertise and meaningful development time no longer holds. When the volume of newly discovered vulnerabilities reaches forty thousand per year and continues rising, patch bandwidth becomes a structural constraint that no organization can solve through effort alone.

The implications for enterprise security architecture are substantial. Patch-centric vulnerability management remains necessary but has become insufficient as the organizing principle of defensive operations. Organizations must supplement it with runtime behavioral monitoring, defense-in-depth architectures designed to limit post-exploitation blast radius, continuous exposure management programs, and AI-assisted prioritization that focuses limited patch bandwidth on the highest-risk items. The central metric for defensive success is shifting from mean time to patch to mean time to detect and contain.

This whitepaper examines the evidence for each of these trends, their interactions and compounding effects, the current state of defensive AI countermeasures, and the strategic adaptations enterprise security programs must make to remain effective in an environment defined by AI-accelerated exploitation.

Introduction: The Vulnerability Window and Why It Is Closing

For most of the history of enterprise vulnerability management, the basic operational model was coherent and defensible. Security researchers or vendors discovered vulnerabilities, coordinated disclosure processes gave vendors time to develop patches, and organizations received notifications that allowed them to plan and execute remediation. Attackers could theoretically exploit vulnerabilities before patches arrived, but doing so required deep technical expertise, significant development effort, and — for widespread exploitation — infrastructure investment that constrained the threat to sophisticated actors. The result was a window of opportunity for defenders: imperfect, variable in duration, and subject to erosion by publicly released proof-of-concept code, but sufficient to sustain a patch-management-centered security posture.

This window emerged from a specific set of technological and economic conditions. Exploit development was a skilled craft requiring detailed understanding of memory layout, operating system internals, and defensive mitigations. Converting a vulnerability disclosure into a reliable exploit could take weeks of skilled engineering work, and weaponizing that exploit for mass deployment required additional infrastructure and operational security investment. These constraints created a time cost for attackers that, while never insurmountable, gave defenders a meaningful operational window — a period during which the probability of encountering a working exploit against a specific unpatched system was low enough that measured remediation programs could manage the risk.

Both the technical constraints and the economic constraints that sustained the patch window are now deteriorating simultaneously. On the technical side, large language models and other AI systems have demonstrated the ability to analyze vulnerability disclosures and generate functional exploit code with substantially reduced human effort. On the economic side, exploit-development-as-a-service platforms and automated vulnerability weaponization pipelines have lowered the cost of deploying working exploits from the domain of nation-state resources to commercially accessible services. And on the discovery side, AI-augmented scanning tools are finding vulnerabilities at a rate that human researchers alone cannot match, compressing the time between when a vulnerability exists in software and when it becomes a publicly known — and publicly exploitable — target.

This paper examines these developments in sequence, building from historical baseline data through current acceleration patterns to strategic recommendations for enterprise security programs. Section 2 presents the longitudinal evidence for declining time-to-exploit over the past several years. Section 3 describes how AI is functioning as an acceleration engine across both offensive and defensive vulnerability operations. Section 4 addresses the specific problem of pre-disclosure exploitation, which is the most structurally destabilizing element of the current threat landscape. Section 5 examines the volume-speed collision that is overwhelming traditional patch management programs. Section 6 reviews the current state of defensive AI

countermeasures and their limitations. Section 7 draws strategic implications for enterprise security architecture and offers recommendations organized by implementation horizon. Section 8 maps these considerations to relevant CSA frameworks and resources.

The Compression of Time-to-Exploit: A Decade of Evidence

The most systematic longitudinal data on exploitation timelines comes from Mandiant, which has tracked the interval between public vulnerability disclosure and first observed exploitation in the wild across thousands of CVEs over multiple years. The data paints an unambiguous picture of accelerating threat.

In 2018 and 2019, the median time-to-exploit across tracked vulnerabilities was sixty-three days — roughly two months during which organizations could realistically identify affected systems, obtain patches from vendors, test compatibility, and deploy remediation. By 2020, that median had fallen to forty-four days. By 2021 and 2022, it had reached thirty-two days, with 44 percent of tracked n-day vulnerabilities exploited within two weeks of disclosure [1]. The 2023 data showed further compression, with Mandiant reporting a median of approximately seven days from disclosure to first publicly available exploit and approximately forty-three days from disclosure to observed first exploitation in the wild — a pattern suggesting that while threat actors are moving faster, they sometimes delay deploying capabilities against hardened targets to maximize operational value [2].

Rapid7's annual vulnerability intelligence reports corroborate this compression and identify cases that are more extreme. In 2022, Rapid7 found that 56 percent of tracked vulnerabilities were exploited within seven days of public disclosure [13]. By 2025, Rapid7 reported that the median time to inclusion on CISA's Known Exploited Vulnerabilities catalog had fallen from 8.5 days to 5 days over the preceding two years, and the mean had dropped from 61 days to 28.5 days, suggesting improvement in both the speed of the fastest attackers and the characteristic response of the broader threat actor population [14]. Quarter-by-quarter data from early 2025 shows that approximately 28 percent of exploits were observed within one day of CVE disclosure — meaning more than one in four exploits now arrive within twenty-four hours of the vulnerability becoming publicly known [14].

The Log4Shell incident provides a widely studied case study in what rapid exploitation looks like at scale. When CVE-2021-44228 became public in December 2021, exploitation attempts were observed within hours, and mass scanning for vulnerable systems began before most organizations had even processed the advisory. The incident forced a reckoning with patch management assumptions: organizations that had built their vulnerability response processes around a multi-week remediation cycle found themselves exposed for days to what was effectively a universally available and trivially weaponized critical flaw. Post-incident

analysis documented that organizations needed to compress remediation from the typical weeks to twelve hours or less just to avoid sustained exposure [15]. The Log4Shell response became a forcing function for organizations to build faster vulnerability management operations, but the subsequent trajectory of time-to-exploit data suggests that the threat has continued to outpace even these improved response programs.

Several factors beyond AI have contributed to this compression over the years — the public availability of proof-of-concept exploit code, bug bounty programs that simultaneously advance disclosure and publish technical details, and the commoditization of offensive security tooling. These factors were well understood and largely built into the operational assumptions of mature vulnerability management programs. What AI introduces is qualitatively different: not an incremental acceleration of existing exploit development processes, but the potential for automated exploit generation at a speed and scale that no human-centered development process can approach.

AI as an Acceleration Engine: Discovery, Development, and Weaponization

AI-Augmented Vulnerability Discovery

The dual nature of AI in vulnerability operations is nowhere more evident than in the evolution of Google's Big Sleep project. Big Sleep is a collaboration between Google Project Zero and Google DeepMind that deploys large language models to autonomously analyze software codebases and identify exploitable vulnerabilities. In October 2024, Big Sleep became the first publicly documented AI system to discover a previously unknown, real-world exploitable vulnerability: a stack buffer underflow in SQLite that was patched the same day by the SQLite development team, before any public release of the vulnerability details [3]. The fact that the patch arrived the same day — before the vulnerability was publicly known — illustrates both the defensive potential and the structural complexity of AI-augmented security research.

The SQLite discovery was significant for reasons beyond the specific vulnerability. Google's researchers noted that this vulnerability was not accessible to traditional fuzzing tools, because the relevant code path required specific configuration conditions that existing OSS-Fuzz harnesses did not set up. Big Sleep found it by reasoning about the code semantically, identifying the logical conditions under which a dangerous state could be reached, rather than by generating random inputs and observing crashes. This semantic reasoning capability — analyzing intent and structure rather than producing inputs by volume — represents a genuinely new paradigm in vulnerability discovery, one that can find entire classes of vulnerabilities that prior automated methods miss.

Since late 2024, Big Sleep has continued to discover vulnerabilities in real-world software at a pace that would be impossible for any individual human researcher or small team. The significance of this trajectory is not only technical. It demonstrates that AI systems operating defensively can find and close vulnerabilities before attackers can exploit them. But the same technical capability that enables defensive discovery is available, in principle, to offensive actors with access to similar systems. An AI system that can find a stack buffer underflow in SQLite by reasoning about code semantics can perform the same analysis on any software to which it has access — including proprietary enterprise software, embedded firmware, or network device code that traditional external security research never reaches.

Google's OSS-Fuzz program, which applies fuzzing techniques to open-source software at scale, has incorporated AI-enhanced components that achieved a 15 percent fix rate for LLM-addressed crash analysis, demonstrating that AI systems are contributing meaningfully to automated remediation rather than merely identification [4]. Research published in late 2024 demonstrated LLM-guided fuzzing systems that discovered substantially more vulnerabilities than prior state-of-the-art approaches on tested codebases, with improvements in both code coverage and semantic diversity of test inputs [5][16]. The ProphetFuzz system used prompt-engineered LLMs to predict vulnerable option combinations and discovered 364 unique vulnerabilities in tested targets — approximately one-third more than the prior leading approach. SyzGPT, applied to kernel fuzzing, reported a 323 percent improvement in vulnerability detection over baseline approaches.

Anthropic's collaboration with Mozilla in early 2026 provides a concrete illustration of what AI-assisted vulnerability discovery looks like at enterprise codebase scale. Scanning approximately six thousand C++ files in the Firefox codebase over roughly two weeks, Claude produced 112 bug reports, of which 22 were confirmed as CVEs and 14 were classified as high-severity [17]. The scale of this result — more than one confirmed CVE per day from a two-week engagement with a single codebase — suggests that AI-augmented security review is finding vulnerabilities at a rate that substantially exceeds what traditional security research engagements produce. Extrapolated across the full landscape of deployed software, the implication is that AI will dramatically increase the total count of known vulnerabilities in coming years, while simultaneously compressing the time between when a vulnerability exists and when it becomes a known target.

AI-Assisted Exploit Development

The same LLM capabilities that enable vulnerability discovery also lower the barrier to exploit development. Research in 2025 demonstrated that AI systems can analyze CVE disclosures and generate functional exploit code with substantially reduced human effort, at an estimated cost of approximately one dollar per exploit attempt [18]. The economic implication is straightforward: exploit development transitions from a skilled craft requiring significant per-target investment to a commodity service with negligible marginal cost.

Estimates suggest that automated pipelines operating at this cost structure can operationalize exploitation capabilities against over one hundred newly disclosed CVEs per day, shifting the attacker's constraint from development capacity to deployment infrastructure and targeting selection [18].

This cost collapse has organizational implications that go beyond individual exploit development. Threat actors who previously had to allocate skilled engineering resources to develop exploits for the highest-value targets can now deploy automation against the full spectrum of disclosed vulnerabilities. The democratization of exploit development means that vulnerability classes previously requiring nation-state-level resources are now accessible to organized criminal groups, and vulnerability classes that previously required advanced criminal groups are accessible to lower-tier actors. The skill distribution of the effective attacker population is shifting downward, while the effective quantity of exploitation attempts directed at any given disclosed vulnerability is shifting upward.

Google Threat Intelligence's analysis of the defensive implications of AI-accelerated vulnerability exploitation identifies the core challenge clearly: defenders must patch every relevant vulnerability within their environment, while attackers need only find one that has not been patched [19]. AI fundamentally advantages the attacker side of this asymmetry by reducing the cost of scanning and exploiting the full vulnerability surface to near zero. The implication is not that patching should be abandoned — it remains the most effective single intervention available — but that the cost-benefit structure of the offense-defense balance has shifted in ways that demand a strategic response beyond simply patching faster.

Pre-Disclosure Exploitation: When the Timeline Inverts

The most structurally disruptive element of the current vulnerability landscape is not accelerated time-to-exploit after disclosure — it is the growing body of evidence that a substantial fraction of vulnerabilities are exploited before they are publicly disclosed at all. This "negative time-to-exploit" scenario is not a theoretical risk; it is a documented and quantifiable pattern in current threat activity.

VulnCheck's analysis of the first half of 2024 identified fifty-three zero-day vulnerabilities with exploitation evidence at or before the date of public disclosure [7]. Miggo Security's research on CISA's Known Exploited Vulnerabilities catalog found that 23.6 percent of tracked KEV entries showed evidence of exploitation on or before their CVE disclosure date [8]. Additional analysis from Zafran's 2024 vulnerability exploitation review found that a substantial portion of exploits were available before CVE numbers were published, and raised the concern that KEV-based prioritization approaches may be systematically under-representing active exploitation activity because only publicly disclosed vulnerabilities can receive CVE designations [11].

The pre-disclosure exploitation pattern has several distinct causes. Vulnerability brokers and government-sponsored offensive programs acquire zero-day vulnerabilities and deploy them covertly, sometimes for months or years before public disclosure occurs. Sophisticated threat actors conducting targeted intrusions discover and exploit vulnerabilities through their own research, with no incentive to notify vendors. In some cases, vulnerability discovery and initial exploitation occur nearly simultaneously — a threat actor scanning for a specific vulnerability class may find and exploit an instance while security researchers are still analyzing the same class. The window between when an attacker understands a vulnerability and when a defender can act on it has, in these cases, effectively collapsed to zero.

AI amplifies each of these patterns. AI-augmented vulnerability discovery enables threat actors to scan large codebases and identify exploitable conditions more quickly than previous research timelines allowed. The speed with which AI can generate exploit code from a vulnerability discovery narrows the window between finding a flaw and deploying a working exploit to hours or less. AI-powered scanning infrastructure can survey the full landscape of internet-facing systems for vulnerable instances faster than disclosure processes can notify affected organizations. The combination creates conditions where the first actor to find a vulnerability — whether a defensive researcher or an offensive one — can move to exploitation faster than any coordinated disclosure process can distribute patches.

The structural implication for enterprise security programs is profound. A vulnerability management program organized around the principle of "patch before exploitation" is operating on an assumption that does not hold for a meaningful fraction of the highest-severity vulnerabilities. CISA's Known Exploited Vulnerabilities catalog is a reactive instrument by design: it documents exploitation that has already been observed and reported. It cannot, by definition, capture exploitation that is occurring without public knowledge. Enterprise security teams that use the KEV catalog as their primary triage signal may be systematically under-responding to the most dangerous threats in their environment, not through any fault of the catalog's design, but because the catalog was designed for a threat landscape where exploitation followed disclosure at a measurable and manageable lag.

This does not mean that patching is unimportant. Patching remains essential, and rapidly patching known exploited vulnerabilities is still the single highest-value preventive action most organizations can undertake. But it does mean that patching cannot be the sole or primary defense against the most sophisticated threats. Organizations must supplement patch management with detection and response capabilities specifically designed to identify post-exploitation behavior, because for a meaningful and growing fraction of vulnerabilities, the exploit will arrive before the patch — or before the vulnerability is even publicly known to exist.

The Volume-Speed Collision

The acceleration of exploitation timelines is compounded by a parallel increase in the absolute volume of new vulnerabilities that organizations must manage. These two trends interact in ways that make the overall challenge substantially more difficult than either would be in isolation.

NIST's National Vulnerability Database recorded 40,009 new CVEs in 2024, an increase of 38 percent over 2023 and an all-time record [21]. The daily rate of new disclosures reached approximately 127 to 131 CVEs per day in 2025, and since 2020, total CVE submissions have increased by approximately 263 percent [20] [21]. The acceleration in discovery rate is itself partially attributable to the same AI-augmented scanning tools described in this paper. As AI systems find more vulnerabilities, the total inventory of publicly known vulnerabilities grows faster, enlarging the denominator of the management problem at precisely the moment when the time available to address each item is shrinking.

Against this backdrop of rising volume, organizational patch deployment capacity has not scaled proportionately. Industry survey data consistently finds that a large majority of organizations require more than one week to deploy patches for critical vulnerabilities, with a significant fraction requiring substantially longer due to testing requirements, vendor dependencies, and legacy system constraints [22]. More than half of security and IT professionals in recent surveys describe the volume of vulnerabilities as exceeding their current capacity to remediate effectively — a finding that represents not a failure of effort but a structural mismatch between the rate of vulnerability production and organizational remediation capacity [22]. It is worth noting that even this picture captures only software vulnerabilities in conventional systems; AI-specific vulnerabilities add a rapidly expanding category that requires different analytical frameworks and is not yet adequately covered by most vulnerability management programs.

AI-specific CVEs represent a particularly challenging dimension of this volume problem. The count of CVEs affecting AI frameworks, inference engines, model serving platforms, and AI-integrated applications grew from approximately three hundred in 2023 to over one thousand in 2025, a compound annual growth rate that substantially exceeds the overall CVE trend [23]. The security properties of AI systems differ meaningfully from conventional software: prompt injection, model poisoning, inference manipulation, and agent hijacking represent attack surfaces with no direct analogues in traditional vulnerability management, requiring security teams already stretched by conventional vulnerability volume to absorb a new and rapidly expanding vulnerability class that requires different tooling and different analytical reasoning.

The combination of rising volume and accelerating exploitation creates a mathematical problem that cannot be solved through increased patching velocity alone. Even an organization that doubles its patch deployment capacity will still face a growing backlog of unpatched vulnerabilities being actively exploited, because exploitation timelines are compressing faster than most organizations can accelerate remediation. NIST's National Vulnerability Database itself experienced analysis and scoring backlogs in 2024 as

disclosure volume exceeded processing capacity, briefly reducing the utility of NVD as a real-time prioritization signal. This strain on even foundational vulnerability infrastructure illustrates the systemic nature of the volume challenge.

CISA's Exploit Prediction Scoring System (EPSS), maintained by the Forum of Incident Response and Security Teams, is the most mature tool currently available for risk-based prioritization under these conditions [6]. EPSS produces a probability score updated daily, estimating the likelihood that a given CVE will be exploited within thirty days, drawing on over a thousand variables that include exploit availability indicators, vulnerability metadata, and historical exploitation patterns. Unlike CVSS, which measures theoretical severity, EPSS measures empirical risk in the current threat landscape. Organizations that triage their vulnerability backlogs using EPSS scores alongside CVSS severity have been shown to substantially improve the efficiency of their patch programs, focusing limited remediation bandwidth on the vulnerabilities that are actually being weaponized rather than those that are theoretically most severe. However, EPSS models are calibrated on historical exploitation patterns that predate the current AI acceleration, and recalibration will be needed to ensure the model continues to accurately reflect a threat landscape where exploitation timelines and volumes are both compressing rapidly.

Defensive AI: Countermeasures and Their Structural Limits

The same AI capabilities that are accelerating offensive exploitation are being actively applied to defensive operations, and the results are substantial. Understanding both the genuine capabilities and the structural limitations of defensive AI is essential for organizations calibrating their strategic response.

Google DeepMind's CodeMender project represents one of the most advanced current applications of AI to vulnerability remediation. CodeMender applies AI-based reasoning not only to identify vulnerabilities but to autonomously generate and contribute security patches upstream to open-source projects. In approximately six months of operation through early 2026, CodeMender contributed seventy-two security fixes to open-source projects spanning over 4.5 million lines of code [24]. This demonstrates that AI can close the loop between discovery and remediation for at least a subset of vulnerability classes, particularly those with well-understood patterns where patch generation is amenable to automated reasoning. Meta's AutoPatchBench, released in April 2025, provides a benchmarking framework for evaluating the effectiveness and reliability of AI-generated security fixes — a necessary precursor to deploying such fixes with confidence in production environments [25].

Microsoft Security Copilot and comparable enterprise security AI platforms are delivering meaningful operational value across large organizations. By integrating vulnerability data with threat intelligence feeds and asset inventory, these tools can substantially reduce the analyst time required to prioritize and triage a large vulnerability backlog. Microsoft has documented use cases where Security Copilot reduces the time required to investigate a vulnerability from hours to minutes, by automating the correlation of CVE data with affected asset information, compensating control status, and threat intelligence context [26]. At the scale of enterprise vulnerability programs managing thousands of outstanding items, these efficiency gains translate into meaningfully faster response for the highest-priority items.

Despite these genuine advances, several structural limitations constrain defensive AI's ability to fully counter the threat landscape described in this paper.

The first constraint is the deployment gap. Identifying a vulnerability and generating a patch is not the same as deploying that patch across an enterprise environment. Enterprise patch deployment involves compatibility testing, managing vendor release cycles, coordinating with change management processes, and handling the substantial fraction of legacy systems for which patches may not be available or for which deployment carries operational risk. AI can dramatically accelerate the front end of the vulnerability management process — identification and prioritization — but the bottleneck in most organizations is in the deployment phase, where organizational and operational constraints dominate in ways that AI cannot currently resolve.

The second structural constraint is the asymmetry of objectives. Offensive AI must succeed once; defensive AI must succeed always. An AI exploit generator that produces a working exploit for 10 percent of targeted vulnerabilities is highly effective in an offensive context because even a low success rate yields a large number of working exploits when applied at machine scale. Defensive AI that correctly identifies 90 percent of exploitable vulnerabilities still leaves a 10 percent residual that represents concrete organizational exposure. This asymmetry is fundamental and cannot be resolved by improving defensive AI capabilities alone, because the failure modes point in opposite directions.

The third structural constraint is the calibration problem. CISA's Known Exploited Vulnerabilities list and EPSS, the two primary tools for empirical vulnerability prioritization, were designed and calibrated for a threat landscape that predates the current AI acceleration. Miggo Security's research found that KEV-based prioritization may substantially underrepresent current exploitation activity because exploitation is now occurring faster than the discovery and reporting chain that populates the catalog [8]. This is not a design flaw in these tools — they represent the best available approach under current conditions — but it does mean that organizations relying on them exclusively may have incomplete visibility into the fraction of their unpatched vulnerabilities that are actively being exploited.

The fourth constraint concerns AI-specific vulnerabilities. As noted above, AI systems introduce vulnerability classes that differ structurally from conventional software vulnerabilities. Defensive AI tools designed for traditional vulnerability management may not adequately address prompt injection, indirect prompt

injection, model extraction, inference manipulation, or multi-modal attack vectors against AI components. The security discipline for AI-specific vulnerabilities is still developing, and the tooling that supports it is correspondingly immature relative to the pace at which AI-integrated systems are being deployed in enterprise environments.

Strategic Implications for Enterprise Security Architecture

The trends described in this paper have concrete implications for how enterprise security programs must be structured. The central implication is that patch-centric vulnerability management, while remaining necessary, cannot be the organizing principle of enterprise defense against the threat landscape that AI is creating. The following strategic adaptations represent the highest-priority areas of investment.

The first and most fundamental shift is from disclosure-triggered to continuous exposure management. The traditional model — monitor for disclosures, assess applicability, schedule remediation — assumes that disclosure is a reliable trigger for the onset of risk. Pre-disclosure exploitation data shows this assumption is flawed for the highest-severity vulnerabilities. Organizations must transition to continuous exposure management programs that maintain an always-current inventory of their external attack surface, internal vulnerability landscape, and compensating control posture. This means treating vulnerability management not as a scheduled maintenance activity but as a continuous security operation that generates actionable intelligence on a real-time basis, informed by threat intelligence feeds that include exploitation-in-the-wild signals rather than just disclosure feeds.

The second strategic shift concerns architecture. As the fraction of vulnerabilities exploited before or immediately upon disclosure increases, the security architecture must be designed to limit post-exploitation impact rather than prevent all initial compromise. Zero trust network architecture, application-layer segmentation, privileged access management, and strong identity controls all reduce the blast radius of a successful exploitation, buying time for detection and response even when prevention fails. An organization with mature zero trust architecture may be meaningfully more resilient than one focused on perimeter defense even when both have identical patch deployment velocity, because the constraint on attacker movement post-exploitation limits the damage that any single unpatched vulnerability can cause.

The third shift involves detection. Exploits developed by AI against recently disclosed or undisclosed vulnerabilities may not match the signatures of known exploit tools. Detection capabilities that focus on behavioral indicators — anomalous process behavior, unexpected network connections, credential usage patterns inconsistent with baselines, lateral movement signatures — are less dependent on prior knowledge

of specific exploit techniques and more effective against novel attack chains. This does not replace signature-based detection but must supplement it as AI-generated exploits become more common and more varied than any signature library can fully anticipate.

Risk-based prioritization using empirical exploitation probability data must replace or substantially augment CVSS-only scoring. Organizations that are not already using EPSS alongside CVSS for patch prioritization are allocating remediation effort to vulnerabilities that are not being exploited while leaving gaps against ones that are. As EPSS models are updated to account for AI-accelerated exploitation patterns, the signal quality will improve. In the interim, organizations should supplement EPSS with active threat intelligence feeds that provide exploitation-in-the-wild signals for the vulnerabilities most relevant to their technology stack.

Software Bills of Materials (SBOMs) have moved from compliance aspiration to operational necessity. Rapid response to vulnerability events like Log4Shell was severely impeded by organizations' inability to quickly determine which of their systems contained the affected component. SBOMs enable fast triage of vulnerability impact across complex software supply chains, and as AI-assisted vulnerability discovery increases the frequency and pace of significant disclosure events, SBOM infrastructure becomes a prerequisite for maintaining acceptable response velocity. Organizations that cannot answer the question "which of our systems uses library X?" within hours of a significant disclosure are structurally unprepared for the current threat environment.

Finally, the primary metrics by which security programs measure and communicate success must shift. Security programs that measure success primarily through patch deployment velocity are measuring the wrong thing in an environment where exploitation may precede patching. Mean time to detect exploitation, mean time to isolate affected systems, and mean time to restore affected services reflect operational resilience in an AI-accelerated threat environment. MTTR and MTTD should be instrumented and reported with the same rigor that organizations currently apply to patch compliance metrics. The organizational conversation about vulnerability management must shift from "how many vulnerabilities have we patched?" to "how quickly can we detect and contain exploitation when it occurs?"

These adaptations are not alternatives to good vulnerability management; they are complements to it that address the portions of the threat landscape that patch management alone cannot reach. Organizations that simultaneously patch known exploited vulnerabilities quickly, maintain continuous exposure management, implement zero trust architecture, deploy behavioral detection, and measure response effectiveness are substantially better positioned than organizations that do any one of these things well in isolation.

CSA Resource Alignment

CSA's AI Safety Initiative and broader security framework portfolio provide directly relevant guidance for organizations navigating the threat landscape described in this paper.

The AI Controls Matrix (AICM) addresses vulnerability management requirements across the AI supply chain, covering model providers, application providers, orchestrated service providers, cloud service providers, and AI customers. As AI-specific CVEs grow to represent a meaningful and rapidly expanding fraction of total vulnerability volume, AICM's controls for model and component security provide a structured framework for addressing vulnerability exposure that conventional vulnerability management programs were not designed to handle. AICM's implementation and auditing guidelines for each stakeholder role offer concrete starting points for organizations building vulnerability management programs for AI-integrated environments. Because AICM is a superset of CCM, organizations already operating under CCM have a foundation on which AI-specific controls can be layered incrementally.

MAESTRO, CSA's threat modeling methodology for agentic AI systems, addresses attack chains that are particularly relevant in an AI-accelerated exploitation context. As enterprises deploy autonomous agents capable of executing multi-step actions across tools, APIs, and data systems, the risk that a compromised or exploited component will be used as a pivot point within an agentic workflow grows substantially. A vulnerability exploited in an AI agent's underlying model or orchestration framework may have consequences that extend well beyond the component itself if the agent has been granted broad tool access. MAESTRO's layer-by-layer threat analysis provides a structured approach to identifying these attack surfaces before they are exploited, and connects directly to the kind of runtime protection and behavioral monitoring that the exploit-before-patch environment demands.

CSA's Zero Trust guidance is directly applicable to the assume-breach architectural shift described above. The principle that no internal network position should be inherently trusted, and that access should be continuously verified based on identity and context rather than network location, becomes substantially more important as exploitation timelines compress and the assumption of clean internal environments becomes less defensible. CSA's publications on communicating the business value of zero trust and CISO perspectives on zero trust deployment provide both the technical framework and the executive communication tools needed to advance zero trust adoption in organizations where leadership still treats it as aspirational rather than urgent.

The Agentic AI Red Teaming Guide, published in 2025, addresses testing methodologies specifically designed for AI-integrated systems. As AI agents become both targets for exploitation and potential vectors within compromised environments, structured red team methodology for these environments is essential preparatory work that most organizations have not yet completed. The exploitation patterns enabled by AI make adversarial testing of agentic systems more, not less, important than testing of conventional applications.

CSA's STAR for AI assurance program provides a framework for organizations to assess and communicate the security posture of their AI systems, including vulnerability management practices specific to AI components. As regulatory expectations around AI security continue to develop globally, and as procurement requirements begin to include AI security assurance criteria, STAR for AI provides a structured evidence framework applicable to demonstrating compliance with emerging requirements.

Conclusions and Recommendations

The exploit-before-patch structural gap is not a future risk that will materialize if AI capabilities continue to develop. It is a present operational reality, documented in longitudinal exploitation data, demonstrated by AI systems that have found real-world vulnerabilities in major software projects, and reflected in the experiences of organizations that have attempted to patch critical vulnerabilities after disclosure only to find that exploitation had already begun.

The core conclusion of this analysis is that the patch-centric vulnerability management model has been structurally undermined by the intersection of AI-accelerated exploit development, pre-disclosure exploitation at measurable scale, and vulnerability volume growth that exceeds organizational remediation capacity. Organizations that continue to treat patch deployment velocity as the primary indicator of vulnerability management health are measuring a necessary condition but not a sufficient one.

The following recommendations are organized by implementation horizon and represent the highest-priority adaptations for enterprise security programs operating in this environment.

Immediate Actions (0–30 Days). Security teams should assess their current dependency on disclosure-triggered patching and determine what fraction of their patch triggers derive from CISA KEV entries, vendor security advisories, or internal vulnerability scanning — and how quickly each of those sources translates to deployed patches. Any organization that cannot achieve critical patch deployment within five to seven days of a KEV listing for their highest-exposure assets should treat this as an immediate operational gap. Security operations teams should also assess whether their detection capabilities are oriented toward behavioral indicators of post-exploitation activity, or primarily toward signatures of known threats; if the latter, a gap assessment against behavioral detection baselines should be initiated. Finally, organizations that are not currently incorporating EPSS scores into vulnerability prioritization should begin that integration without delay.

Short-Term Adaptations (30–90 Days). Organizations should evaluate SBOM coverage across their software supply chain and develop a roadmap for achieving the component visibility needed to conduct rapid triage of major vulnerability events. The capacity to answer "which of our systems uses this library?" within hours of a major disclosure should be treated as a minimum operational requirement, not an

aspirational goal. For AI-integrated environments specifically, a dedicated review of AI-component vulnerability management should be conducted using AICM as a baseline framework, with particular attention to AI-specific vulnerability classes — prompt injection, model-layer attacks, agent hijacking — that conventional vulnerability management programs do not address. Security teams should also run tabletop exercises that assume initial access has already occurred through an unpatched or pre-disclosure vulnerability, and test detection and response paths under that assumption.

Strategic Investments (3–12 Months). The strategic program required to address the structural gap described in this paper involves four parallel workstreams. First, maturing continuous exposure management to replace periodic patch cycles with real-time visibility into exposure and exploitation status. Second, advancing zero trust architecture to reduce blast radius of successful exploitation, with particular emphasis on privileged access management and identity-based segmentation that limits lateral movement even when a perimeter has been breached. Third, developing detection and response capabilities specifically oriented toward AI-generated and AI-assisted attack chains, including behavioral analytics baselines for high-value systems that can detect novel exploitation patterns not captured in signature databases. Fourth, building the organizational metrics and reporting infrastructure needed to measure and communicate security posture in terms of detection and response effectiveness rather than patch compliance alone.

The AI acceleration of vulnerability exploitation represents a permanent structural shift in the threat landscape, not a temporary spike that will normalize as defenses catch up. The pace at which AI systems are discovering vulnerabilities, generating exploits, and enabling pre-disclosure exploitation is increasing as the underlying technology improves. Organizations that adapt their security architecture to this new baseline — supplementing patch management with the detection, architectural, and operational capabilities described in this paper — will be substantially more resilient. The window for making these adaptations before experiencing their absence as an operational emergency is present but narrowing.

References

[1] Mandiant / Google Cloud. "Time-to-Exploit Trends: 2021–2022." Google Cloud Threat Intelligence Blog, May 2023.

[2] Mandiant / Google Cloud. "Time-to-Exploit Trends: 2023." Google Cloud Threat Intelligence Blog, February 2024.

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

[4] Google Security Blog. "Leveling Up Fuzzing: Finding More Vulnerabilities with AI." Google Security Blog, November 2024.

[5] Kang, Myeongsoo et al. "Hybrid Fuzzing with LLM-Guided Input Mutation." arXiv preprint arXiv:2511.03995, November 2024.

[6] FIRST (Forum of Incident Response and Security Teams). "Exploit Prediction Scoring System (EPSS)." FIRST.org, 2024.

[7] VulnCheck. "State of Exploitation: 1H 2024." VulnCheck Research, July 2024.

[8] Miggo Security. "Report: Missing 88% of Exploits — Rethinking KEV in the AI Era." Miggo Research, 2025.

[9] CISA. "Known Exploited Vulnerabilities Catalog." Cybersecurity and Infrastructure Security Agency, 2024.

[10] Rapid7. "2025 Vulnerability Intelligence Report." Rapid7 Research, 2025.

[11] Zafran. "Vulnerability Exploitation in 2024." Zafran Security, January 2025.

[12] CERT-EU. "AI Is Changing the Economics of Vulnerability Discovery — Defenders Must Adapt." CERT-EU Blog, 2025.

[13] Rapid7. "2022 Vulnerability Intelligence Report." Rapid7 Research, 2022.

[14] Saptang Labs. "From 48 Hours to Minutes: Why Time-to-Exploit Is Shrinking Faster Than Patch Cycles." Saptang Labs, 2025.

[15] Trull, Jonathan. "Remediating Critical Vulnerabilities in 12 Hours or Less: Lessons Learned from Log4j." Cloud Security Alliance / Qualys, 2022.

[16] Deng, Yinlin et al. "Large Language Models for Fuzzing: All You Need Is a Fuzzing Brain." arXiv preprint arXiv:2509.07225, 2025.

[17] Anthropic. "Using Claude for Security Research: Firefox Vulnerability Discovery." Anthropic, 2026.

[18] Integration Security. "In the Age of AI: The Vanishing Gap Between Vulnerability Disclosure and Exploitation." Integsec Blog, 2025.

[19] Google Cloud Threat Intelligence. "Defending the Enterprise When AI Finds Vulnerabilities Faster." Google Cloud Blog, 2025.

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

[21] NIST. "National Vulnerability Database." National Vulnerability Database, 2025.

[22] DevOps.com. "Patch or Perish: The Brutal Truth About Vulnerability Management in 2025." DevOps.com, 2025.

[23] MITRE / CVE Program. "CVE and AI-Related Vulnerabilities." CVE Blog, July 2024.

[24] Google DeepMind. "Introducing CodeMender: An AI Agent for Code Security." Google DeepMind Blog, October 2025.

[25] Meta AI Research. "AutoPatchBench: Benchmarking AI-Powered Security Fixes." Meta AI, April 2025.

[26] Microsoft. "Microsoft Security Copilot." Microsoft Learn, 2025.