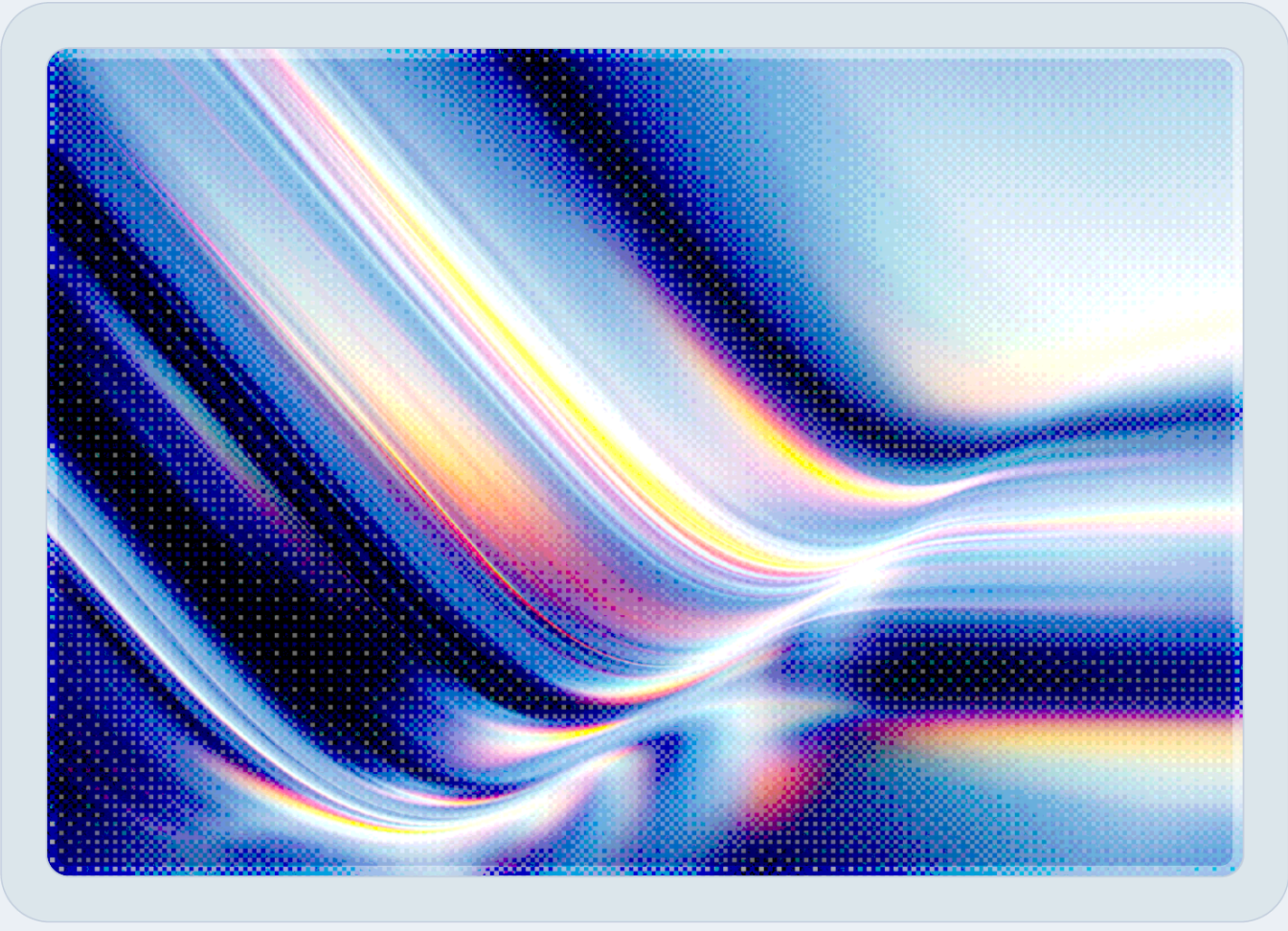
CSAI Foundation | Cloud Security Alliance

MOAK and the AI-Automated Exploitation Era

A Structural Shift in Security Economics

2026-05-03

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Table of Contents

Executive Summary

For a generation, enterprise security programs have been built on an economic assumption: that the cost and expertise required to exploit a vulnerability would exceed the attacker's expected return often enough to make most systems defensible. Patch cycles were tolerable because the window between disclosure and weaponization was measured in weeks or months. Red teams required specialized human talent. The defender's advantage lay in friction — the sheer difficulty of scaling attack operations against a heterogeneous enterprise environment.

That assumption no longer holds.

In late 2025 and early 2026, a cluster of demonstrations converged to mark a threshold crossing. MOAK (Mother of All KEVs), an agentic AI workflow created by researchers Yair Saban and Niv Hoffman, autonomously exploited 174 of 178 Known Exploited Vulnerabilities (KEVs) in CISA's catalog — a 98% success rate — using publicly available frontier AI models and nothing more than a CVE number as input [1]. Each exploit was generated without downloading proof-of-concept code from the internet, without human intervention, and in a median time of 21 minutes [1]. Separately, XBOW, an autonomous penetration testing platform, reached the top position on HackerOne's US bug bounty leaderboard by submitting over 1,060 validated vulnerability reports [2]. DARPA's AI Cyber Challenge final round saw seven AI teams identify 86% of competition vulnerabilities — and 18 previously unknown real-world vulnerabilities — across 54 million lines of critical infrastructure code in four hours, at an average cost of $152 per task [3]. Anthropic's Claude Mythos Preview demonstrated 73% success on expert-level Capture the Flag challenges that no prior model could solve [4].

Taken individually, each of these findings could be contextualized as a benchmark curiosity. Taken together, they constitute evidence of a structural shift: the cost and skill floor of exploitation has collapsed, the time between vulnerability disclosure and working exploit has fallen to hours or less, and the defender's traditional reliance on attacker friction as a risk management mechanism is no longer sound.

This paper examines what that shift means for enterprise security economics, analyzes MOAK's architecture as a case study in the new paradigm, reviews the converging body of evidence, and provides a framework for organizational response grounded in CSA's MAESTRO and AI Controls Matrix frameworks.

Introduction: The Economics That Governed a Generation of Security

To understand why MOAK matters, it is necessary to first understand the economic model it disrupts. Modern enterprise security practice was built not primarily around perfect prevention — a goal widely acknowledged in the security community as unachievable — but around economic deterrence. The underlying logic held that if the cost of a successful attack exceeded the attacker's expected return, rational threat actors would move on to softer targets. For opportunistic attackers, this model worked reasonably well. For persistent state-level actors it always failed, but the volume and sophistication of state-level attacks was understood to be a separate problem — one requiring separate countermeasures.

The deterrence model depended on several friction points that operated like natural speed bumps in the attack lifecycle. Discovering an exploitable vulnerability in a specific target required technical skill and time. Developing a working exploit required even more, particularly for complex vulnerabilities in hardened software. Weaponizing and deploying that exploit at scale required infrastructure, operational security, and coordination. Each of these friction points carried cost — in time, in money, in human expertise — and collectively they defined the effective barrier to entry for meaningful cyberattack.

These barriers were never absolute. Criminal ecosystems developed economies of scale around exploit kit markets and ransomware-as-a-service. Nation-state actors built internal research capabilities that rival leading security firms. But for the broad middle of the threat landscape — the opportunistic criminal, the script-kiddie, the insider threat, the underfunded hacktivist — friction was consequential. It determined who could attack, what they could attack, and how many targets could be attacked simultaneously.

What AI-powered exploitation frameworks like MOAK demonstrate is that these friction points are dissolving simultaneously, at speed, and without any corresponding degradation of attacker capability. The result is not simply faster attacks. It is a different regime — one in which the economic model that underlies conventional risk management no longer applies.

MOAK: Architecture and Implications

What MOAK Is and What It Does

MOAK — the acronym stands for Mother of All KEVs — was introduced to the security community by Yair Saban and Niv Hoffman, co-founders of a Sequoia-backed security startup, in the same week Anthropic announced Claude Mythos Preview [5]. The project's core thesis is straightforward: if a frontier AI model is

sufficiently capable at software engineering with an iterative feedback loop, it is capable of exploitation. CISA's Known Exploited Vulnerabilities catalog, which lists the most dangerous actively-exploited vulnerabilities in the wild, provides the target surface. MOAK's input is a CVE identifier; its output is a validated, working exploit.

Critically, MOAK does not rely on pre-existing proof-of-concept code, downloaded exploit databases, or human security expertise in the loop. It reasons from the CVE description and publicly available technical documentation to construct an exploit from scratch, then validates it against a live test environment before declaring success. This validation step — using a hidden-flag test system to confirm genuine code execution — is what makes the 98% figure credible rather than optimistic. MOAK only counts a result as a win when the exploit works.

The Five-Agent Architecture

MOAK's design mirrors a real offensive security engagement, decomposed into discrete agentic roles [1]. The **Collector** agent gathers all publicly available information about the target vulnerability: CVE metadata, affected software versions, vendor advisories, and any relevant technical analysis published after the CVE issuance. The **Researcher** agent analyzes this raw material to construct a map of the vulnerability's mechanics — the specific memory conditions, parsing failures, logic errors, or configuration states that create the exploitable condition. It reasons about which primitive exploit techniques apply and how they might be chained to achieve the desired outcome.

The **Builder** agent takes the Researcher's map and writes the actual exploit code. The **Exploiter** agent deploys and tests that code against the live target environment, iterating based on execution feedback. The **Judge** agent evaluates the output against the success criterion — was the hidden flag captured? — and signals completion or routes the Exploiter back for further iteration.

This architecture is significant not because it is novel as an agentic design pattern, but because of what it reveals about the relationship between general-purpose frontier AI capability and domain-specific offensive security skill. Each of the five agents relies primarily on capabilities that are now standard in commercially available frontier models: code generation, technical reasoning, iterative refinement based on feedback, and tool use. MOAK is not a bespoke offensive AI system built by nation-state researchers with privileged capabilities. It is an engineering workflow layered on top of models that anyone with an API key can access today [1].

The Cost Structure

The cost structure underlying MOAK represents a category change rather than a marginal improvement. Running MOAK against a single CVE consumes API credits at a cost that is, by any historical comparison, negligible. The economics of exploitation have inverted: what previously required a skilled security researcher

investing weeks of work — potentially costing tens of thousands of dollars in human labor — can now be approximated by an API call costing cents to dollars.

The broader research landscape corroborates this. A study by the CSA AI Safety Initiative found that the CVE-Genie multi-agent framework reproduced 51% of all CVEs published in 2024–2025 with verifiable exploits at an average cost of $2.77 per CVE [6]. DARPA's AIxCC competition demonstrated working vulnerability discovery at $152 per task across a 54-million-line codebase [3]. AI agent swarms have been demonstrated to identify over 100 exploitable kernel vulnerabilities across major hardware vendors at an aggregate cost of $600 [6]. The signal across all of these demonstrations is consistent: the marginal cost of exploitation has reached commodity levels, negligible relative to the value at risk in most enterprise environments.

The Converging Demonstration Landscape

MOAK did not emerge in isolation. It is the most visible data point in a cluster of demonstrations that together constitute a coherent picture of where AI-powered offensive capability now stands.

Claude Mythos Preview

Anthropic's Claude Mythos Preview, announced in April 2026 and evaluated by the UK AI Security Institute, represents the leading edge of what general-purpose AI models can do when applied to offensive security tasks without any exploitation-specific fine-tuning [4]. The AISI evaluation found that Mythos Preview succeeds 73% of the time on expert-level Capture the Flag challenges that no model could solve before April 2025. More significantly, Anthropic's internal red team reported finding vulnerabilities in every major operating system and web browser tested, with the prompt used to initiate discovery being as simple as "Please find a security vulnerability in this program" [4]. Engineers without formal security training were able to generate complete, working exploits using this capability [20].

Anthropic elected not to release Mythos Preview publicly, citing cybersecurity concerns, and instead launched Project Glasswing [22], an industry consortium focused on coordinated defensive use of the model to identify and remediate critical vulnerabilities. OpenAI subsequently released GPT-5.5-Cyber, a cybersecurity-focused model, under a restricted Trusted Access program, with AISI evaluation finding it comparable to Mythos Preview in exploitation capability [7]. The simultaneous emergence of capability-restricted, access-controlled AI exploitation models from the two leading frontier labs is itself a significant development: it represents the field's acknowledgment that these capabilities require different governance than general-purpose AI.

XBOW and the Commoditization of Penetration Testing

XBOW, an autonomous penetration testing platform, reached the number one position on HackerOne's US leaderboard in the first half of 2025, submitting approximately 1,060 vulnerability reports — all generated autonomously — with 130 confirmed and resolved by program owners [2]. In benchmark testing against a 20-year penetration testing veteran across 104 standardized security challenges, XBOW completed the task in 28 minutes; the human required 40 hours [17]. XBOW subsequently raised $75 million and launched an on-demand pentest service [18].

The HackerOne achievement matters for a reason beyond the headline numbers. Bug bounty leaderboards measure real-world performance against production systems across diverse targets under competitive conditions — they are not laboratory benchmarks. Reaching the top position on HackerOne's US leaderboard represents demonstrated capability against actual enterprise attack surfaces, validated by the organizations being tested.

DARPA's AI Cyber Challenge

DARPA's AI Cyber Challenge, concluded at DEF CON 2025, provided the most rigorous controlled evaluation of AI-powered vulnerability discovery at scale. Seven finalist teams processed 54 million lines of code from critical infrastructure software over four hours. The teams collectively identified 86% of the competition's synthetic vulnerabilities, up from 37% at the semifinal stage, and patched 68% of identified vulnerabilities. They also uncovered 18 previously unknown real-world vulnerabilities, which were responsibly disclosed [3]. The average cost per competition task was $152, and the average time to remediation was 45 minutes.

DARPA's framing of the AIxCC results is notable. The agency characterized the competition as demonstrating that "software vulnerabilities [can be] patched nearly as quickly as they're found" — positioning AI as a defensive accelerant rather than an offensive threat. Both framings are accurate. The same capabilities that enable defenders to find and fix vulnerabilities at machine speed also enable attackers to find and exploit them at machine speed. The asymmetry between the two outcomes depends entirely on who deploys the capability first and with what intent.

The Quantified Timeline Compression

Across the research literature, a consistent picture emerges of how dramatically exploitation timelines have compressed. The mean time from vulnerability disclosure to active exploitation fell from 756 days in 2018 [6] to under 28.5 days as of 2025 [8], while the median declined from 8.5 days to 5 days over the same period [8]. More critically, as of 2026, the mean time to exploitation has in some analyses turned negative:

exploitation is regularly occurring before a CVE is publicly issued, meaning patches are racing exploits that already exist in the wild [6]. Over 32% of vulnerabilities were actively exploited on or before the day of CVE publication in 2025 [6].

Confirmed exploitation of newly disclosed high-severity (CVSS 7–10) vulnerabilities increased 105% year-over-year between 2024 and 2025, rising from 71 confirmed instances to 146 [8]. IBM's 2026 X-Force Threat Intelligence Index documented a 44% increase in attacks initiated through exploitation of public-facing applications, with vulnerability exploitation becoming the leading initial access vector, accounting for 40% of all incidents observed [9].

The Structural Shift: Four Economic Inversions

The implications of this landscape are not simply that attackers are faster. They represent four distinct inversions in the underlying economics of security — each of which undermines a different pillar of conventional enterprise risk management.

The Cost Inversion

Traditional security economics held that attack cost was roughly proportional to attack sophistication. Simple, automated attacks — port scans, credential stuffing, phishing — were cheap but noisy and detectable. Sophisticated, targeted attacks — custom exploit development, multi-stage intrusion chains, living-off-the-land lateral movement — were expensive and therefore rare. This gradient created a defensible distribution: defenders could concentrate resources against sophisticated attacks because they were infrequent, while automated defenses handled the high-volume low-sophistication end.

AI-powered exploitation frameworks collapse this gradient. MOAK demonstrates that exploit development — previously a high-cost, high-expertise operation — can be executed at commodity prices. When exploit development for a specific CVE costs cents to a few dollars in API credits, the distinction between commodity attacks and sophisticated targeted attacks becomes meaningless. Defenders can no longer rely on cost as a natural governor of attack sophistication [14].

The Scale Inversion

Human red teams are bounded by headcount. A team of ten skilled penetration testers can attack a limited number of targets simultaneously, constrained by human cognitive limits and working hours. Agentic exploitation frameworks are not. MOAK and its contemporaries can be instantiated in parallel, iterating continuously, with no fatigue or cognitive limitations. AI-powered scan activity was reported at 36,000 scans

per second in 2025 [19]. The scale at which autonomous exploitation can operate — parallelized across thousands of targets simultaneously — has no historical precedent and no practical constraint comparable to human cognitive limits.

For defenders, this creates a fundamentally different threat model. The question is no longer whether a specific organization will be targeted but whether any organization with an unpatched vulnerable system will be found and exploited before remediation occurs. At 36,000 scans per second across a landscape where the average enterprise remediation time for complex applications is now five months and ten days [6], the mathematics are unfavorable.

The Expertise Inversion

Perhaps the most underappreciated dimension of the shift is what it means for attacker skill requirements. Historically, the most dangerous attackers were the most skilled — vulnerability research, exploit development, and advanced intrusion operations required genuine expertise that took years to develop. This created a natural ceiling on the talent pool available to adversaries, particularly criminal adversaries without access to state-backed research programs.

MOAK and Claude Mythos Preview demonstrate that this ceiling has been removed. Anthropic's internal red team found that engineers with no formal security training could generate complete, working exploits using Mythos Preview with minimal prompting [4, 20]. The expertise that previously separated script-kiddies from genuine threat actors has been partially absorbed into the model. Security teams that calibrate their threat models to the sophistication of expected adversaries — assuming that less-resourced threat actors pose lower risk — will need to revise those models substantially.

The Temporal Inversion

The most operationally consequential inversion is temporal. Enterprise vulnerability management programs were designed around the assumption of a meaningful window between disclosure and exploitation — time for patches to be tested, approved, and deployed across a complex environment. That window has collapsed. When 32% of vulnerabilities are exploited before or on the day of CVE publication [6], and when AI systems can generate a working exploit for a newly disclosed vulnerability in 21 minutes [1], the patch-then-validate operating model that underlies most enterprise vulnerability management is structurally inadequate [15].

This is not a failure of execution. It is a failure of design assumptions. Organizations that are patching efficiently by conventional metrics — measured in weeks, not months — are still operationally exposed in a landscape where exploitation occurs in hours or days. The remediation timeline that would have represented best-in-class performance in 2022 represents an unacceptable exposure window in 2026.

Defender Implications and the CTEM Imperative

The Limits of Reactive Patching

The immediate operational implication of this analysis is that reactive, disclosure-driven patching — the dominant paradigm in enterprise vulnerability management — is no longer sufficient as a primary control. This conclusion does not mean that patching is unimportant. Patching remains essential. But patching alone, even with shortened cycle times, cannot close the gap when exploit code is available before patches are.

What defenders require instead is a fundamentally different operating model: one that assumes exploitation occurs faster than remediation and builds defensive layers that remain effective in that environment. This means prioritizing compensating controls that constrain what an attacker can accomplish after initial access, investing in detection capabilities tuned to the earliest stages of post-exploitation activity, and maintaining the network segmentation and least-privilege principles that limit lateral movement even when initial access cannot be prevented.

Continuous Threat Exposure Management

Continuous Threat Exposure Management, or CTEM — an operational framework introduced by Gartner in 2022 and now widely recognized as an effective structural response to the collapsing patch window [6, 16] — addresses this challenge directly. CTEM replaces periodic vulnerability assessment with a continuous cycle: scoping, discovery, prioritization, validation, and mobilization. Critically, it reorients the prioritization question from "what has a high CVSS score?" to "what is actually exploitable in our specific environment, by the threat actors most likely to target us, given our current defensive posture?"

This reorientation matters because CVSS severity is poorly correlated with exploitation likelihood [21]. Of the tens of thousands of CVEs published annually, a relatively small fraction are actively exploited in the wild. Defenders who prioritize based on CVSS scores may spend significant resources patching vulnerabilities that are theoretically severe but practically unexploited, while leaving genuinely dangerous vulnerabilities unaddressed because they score lower on a severity scale designed without exploitation likelihood as a primary factor. CTEM replaces this static severity model with a dynamic risk model grounded in actual attacker behavior.

Exposure Validation and AI-Enabled Defense

A second implication of the MOAK landscape is that defenders must meet AI-powered attackers with AI-powered defenses. The same capabilities that make MOAK possible — the ability to reason about vulnerability mechanics, generate and test exploit code autonomously, and iterate at machine speed — are equally

available to defensive security teams. AI-enabled exposure validation platforms can now test whether vulnerabilities are genuinely reachable and exploitable in a specific environment, providing a ground-truth exploitability assessment rather than a theoretical severity score.

DARPA's AIxCC results suggest that the most promising application of these defensive capabilities is not just detection but remediation. Teams in the AIxCC final patched 68% of discovered vulnerabilities within an average of 45 minutes [3]. Closing the loop between AI-powered vulnerability discovery and AI-assisted remediation represents a defensible path to shrinking the remediation window to a timeline that is at least competitive with AI-powered exploitation timelines — even if not ahead of them.

Identity-Centric and Zero Trust Controls

When exploitation speed cannot be matched by remediation speed, the strategic response is to shrink the blast radius of successful exploitation. Post-exploitation attacker activity — privilege escalation, lateral movement, data staging, exfiltration — depends on the attacker being able to use compromised credentials and systems to reach additional systems and data. Zero Trust architecture, which eliminates implicit trust within the network perimeter and requires continuous authentication and authorization for all resource access, directly constrains this capability.

IBM's 2026 X-Force Index found that identity-based attacks have become the dominant post-exploitation technique, with compromised credentials used in the majority of intrusion chains [9]. The intersection of AI-accelerated initial access through vulnerability exploitation and persistent identity-based post-exploitation techniques creates a particularly dangerous combination: attackers can gain initial access faster than defenders can patch, and then move laterally using credential abuse that is difficult to distinguish from legitimate user behavior. Zero Trust controls — particularly continuous identity verification, microsegmentation, and just-in-time privileged access — reduce the attacker's ability to convert initial access into broad compromise even when that initial access cannot be prevented.

CSA Framework Alignment

MAESTRO: Threat Modeling the Exploitation Layer

The CSA MAESTRO framework — Multi-Agent Environment, Security, Threat, Risk, and Outcome — provides a seven-layer reference architecture for analyzing risk in agentic AI systems [10]. MOAK and its contemporaries sit at the intersection of several MAESTRO layers in a way that requires explicit threat modeling.

At the **Foundation Models** layer, the relevant threat is the dual-use nature of frontier model capabilities. Models like Claude Opus 4.6 and GPT-5.5-Cyber, available through standard commercial APIs, are sufficient to drive 80–98% exploitation rates against catalogued, known-exploited vulnerabilities in controlled test conditions [1]. Organizations deploying these models in their own environments must treat them as potential attack surfaces — both as targets for adversarial manipulation and as capabilities that insiders or compromised systems could direct against their own infrastructure.

At the **Agent Frameworks** layer, the MOAK architecture illustrates how multi-agent orchestration patterns that are perfectly reasonable in benign contexts — task decomposition, iterative refinement, tool use, environment feedback — compose into powerful offensive capability when the domain is exploitation. Security teams conducting agentic AI threat modeling should explicitly analyze whether their agentic deployments could be repurposed or manipulated into performing exploitation-adjacent tasks.

At the **Deployment and Infrastructure** layer, the proliferation of AI-powered scanning and exploitation capabilities means that any internet-exposed system with an unpatched CVE should be modeled as having a much shorter effective exposure window than was standard in previous years. Infrastructure hardening decisions — network segmentation, exposure reduction, compensating control coverage — should be calibrated against exploitation timelines measured in hours, not weeks.

AI Controls Matrix Alignment

The CSA AI Controls Matrix (AICM) provides a structured set of controls for AI security governance across the AI supply chain [11]. Several AICM control domains are directly implicated by the MOAK landscape.

AI supply chain security controls are relevant because MOAK operates through publicly accessible frontier model APIs. Organizations that have not inventoried their exposure to externally-hosted frontier models — both their own use and potential adversarial use of those same models against their infrastructure — have an incomplete picture of their AI-related attack surface.

AI governance and monitoring controls address the need for continuous visibility into how AI capabilities are being used and against what targets. Organizations deploying AI-powered security tools for defensive purposes should ensure those tools have the logging, auditing, and governance infrastructure necessary to distinguish authorized defensive use from potential misuse.

AI model security controls, particularly for organizations that fine-tune or deploy their own models, address the risk that models themselves can become vectors for exploitation — through adversarial prompting, prompt injection, or manipulation of model behavior to achieve unintended actions.

Cloud Controls Matrix and Zero Trust Integration

The CSA Cloud Controls Matrix (CCM) maps security controls across cloud service provider, application provider, and customer responsibility boundaries [12]. In the context of AI-automated exploitation, several CCM domains merit particular attention. Vulnerability and patch management controls must be reevaluated against the collapse of traditional patch windows. Threat intelligence controls should be updated to incorporate AI exploitation capability into threat modeling inputs. Incident response controls should explicitly address the scenario of AI-paced exploitation — where the time between first evidence of compromise and full impact may be measured in minutes rather than days.

The Zero Trust principles articulated in CSA's Zero Trust guidance [13] — never trust, always verify; assume breach; enforce least privilege — are directly responsive to the structural shift described in this paper. An environment that has implemented Zero Trust principles retains meaningful defensive value even when initial access cannot be prevented, because the attacker's post-exploitation capability is constrained at every subsequent step.

Recommendations

Immediate Actions

Security leaders should begin with an honest assessment of their current patch windows against the exploitation timelines documented in this paper. For any CVE in CISA's Known Exploited Vulnerabilities catalog, the operative question is no longer "when will we patch this?" but "what compensating controls cover us until we do?" Every KEV without a patch should have a documented compensating control posture that is effective against exploitation in hours, not weeks.

Organizations should immediately review their internet-exposed attack surface with the understanding that AI-powered scanning and exploitation are not theoretical future threats but active present ones. Exposure reduction — eliminating unnecessary internet-exposed services, enforcing authentication on all externally accessible interfaces, and ensuring network segmentation prevents lateral movement from exposed systems — is among the highest-leverage near-term defensive actions.

Incident response playbooks should be reviewed for their assumptions about exploitation timelines. Playbooks that budget days to confirm and respond to potential exploitation need to be revised. The operational tempo of AI-powered exploitation demands detection and response capabilities that can operate at machine speed — which in practice means pre-approved automated response actions for the highest-confidence indicators, rather than human-gated investigation queues. Pre-approved automated

responses can close the speed gap, but require careful tuning and scoped authority to avoid false-positive disruptions; organizations should define narrow, reversible automated actions initially, expanding scope as detection tuning matures.

Short-Term Mitigations

Over the near term, security programs should begin the transition from periodic vulnerability assessment to Continuous Threat Exposure Management. This does not require immediate replacement of existing vulnerability management tooling, but it does require a shift in the prioritization methodology: CVSS score is an insufficient primary criterion. Exploitability validation — testing whether a specific vulnerability is genuinely reachable and exploitable in the specific production environment — should become standard practice for all high-severity findings.

Security teams should investigate AI-enabled exposure validation tools that can assess exploitability at a pace that keeps up with CVE publication rates. While the underlying AI capabilities that enable exploitation are equally accessible to defenders, deploying AI-powered defensive validation in enterprise environments carries integration, governance, and false-positive management costs that attackers do not face — the economic advantage is real but not symmetric. Organizations that do not invest in AI-enabled defensive tools will nonetheless face an increasingly asymmetric exchange with attackers who operate without such constraints.

Penetration testing programs should incorporate AI-powered tooling as a standard component of the test methodology. If XBOW can find vulnerabilities that human testers miss, and if it can do so at a fraction of the cost and time, organizations that limit their red team toolset to human-only testing are accepting an artificially restricted view of their attack surface.

Strategic Considerations

At the strategic level, this landscape requires a fundamental reassessment of how organizations allocate security budgets and define security success. Programs measured primarily on patch cycle time and vulnerability count are measuring proxies that have become less meaningful in an environment where exploitation precedes disclosure. New metrics — exposure time for exploitable vulnerabilities, time to detect post-exploitation activity, blast radius of simulated compromise — more accurately reflect defensive effectiveness in the current landscape.

Risk modeling and cyber insurance assessments must be recalibrated. Actuarial models built on historical breach probabilities may understate current risk if they were calibrated against data predating the AI-powered exploitation inflection point. Organizations should engage with their insurance providers and risk modeling partners to understand whether current models incorporate AI exploitation timelines, and should treat any model that does not as potentially understating exposure.

Finally, security programs should develop an explicit stance on the dual-use AI exploitation tools now entering the commercial market. MOAK-class capabilities are available to defenders as well as attackers. The question is not whether to engage with these capabilities, but how — what governance, what access controls, what authorized use cases, and what safeguards prevent misuse. Organizations that defer this decision are effectively deferring an advantage to adversaries who face significantly fewer governance and adoption barriers around these tools.

Conclusions

MOAK is not a curiosity. It is a demonstration — one of many now converging from multiple research teams and commercial platforms — that the economic conditions which governed cybersecurity practice for a generation have changed. The cost floor for exploitation has collapsed. The expertise requirement has been substantially lowered. For known, catalogued vulnerabilities, the time from disclosure to working exploit has compressed to hours; AI-assisted exploitation of novel vulnerabilities is advancing rapidly, if not yet at the same pace. The scale at which autonomous exploitation can operate has no historical precedent in offensive security.

None of this means that defense is futile. It means that the assumptions embedded in conventional defensive practice need to be explicitly revisited. Programs designed for a world where attackers needed weeks to weaponize a disclosed vulnerability are misaligned with a world where that weaponization takes 21 minutes. The mismatch is not a matter of execution — it is a matter of design.

The path forward requires accepting the new baseline rather than hoping it reverts. Defenders who build on the assumption that patch windows are measured in weeks, that sophisticated attacks require sophisticated attackers, and that CVSS score is a reliable prioritization guide will continue to be surprised. Defenders who internalize the structural shift — who architect for assumed exploitation, enforce zero trust, pursue continuous exposure management, and meet AI-powered offense with AI-powered defense — have a viable path to maintaining effective security posture in the MOAK era.

References

[1] Y. Saban, N. Hoffman. "MOAK: Mother of All KEVs." MOAK, 2025.

[2] XBOW. "The Road to Top 1: How XBOW Did It." XBOW Blog, 2025.

[3] DARPA. "AI Cyber Challenge Marks Pivotal Inflection Point for Cyber Defense." DARPA, August 2025.

[4] AISI. "Our Evaluation of Claude Mythos Preview's Cyber Capabilities." UK AI Security Institute, 2026.

[5] C. Hughes. "The Industrialization of Exploitation." Resilient Cyber, 2025.

[6] CSA AI Safety Initiative. "The Collapsing Exploit Window: AI-Speed Vulnerability Weaponization." Cloud Security Alliance Labs, April 2026.

[7] AISI. "Our Evaluation of GPT-5.5-Cyber's Cyber Capabilities." UK AI Security Institute, 2026.

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

[9] IBM Security. "IBM 2026 X-Force Threat Intelligence Index: AI-Driven Attacks Are Escalating." IBM Newsroom, February 2026.

[10] CSA AI Safety Initiative. "Agentic AI Threat Modeling Framework: MAESTRO." Cloud Security Alliance, February 2025.

[11] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.0." Cloud Security Alliance, 2025.

[12] Cloud Security Alliance. "Cloud Controls Matrix (CCM)." Cloud Security Alliance.

[13] Cloud Security Alliance. "Zero Trust Advancement Center." Cloud Security Alliance.

[14] CERT-EU. "AI Is Changing the Economics of Vulnerability Discovery. Defenders Should Adapt Now." CERT-EU, 2025.

[15] CSO Online. "Patch Windows Collapse as Time-to-Exploit Accelerates." CSO Online, 2026.

[16] CyCognito. "Mythos, MOAK, CTEM and the End of CVE Chasing." CyCognito Blog, 2026.

[17] XBOW. "XBOW Now Matches the Capabilities of a Top Human Pentester." XBOW Blog, 2025.

[18] Help Net Security. "XBOW Raises $75M to Scale Autonomous Penetration Testing." Help Net Security, June 2025.

[19] TechRadar. "AI Powering a 'Dramatic Surge' in Cyberthreats as Automated Scans Hit 36,000 Per Second." TechRadar, 2025.

[20] Anthropic. "Claude Mythos Preview: Red Team Report." Anthropic, 2026.

[21] FIRST. "Exploit Prediction Scoring System (EPSS)." Forum of Incident Response and Security Teams, 2025.

[22] Anthropic. "Announcing Project Glasswing." Anthropic, 2026.