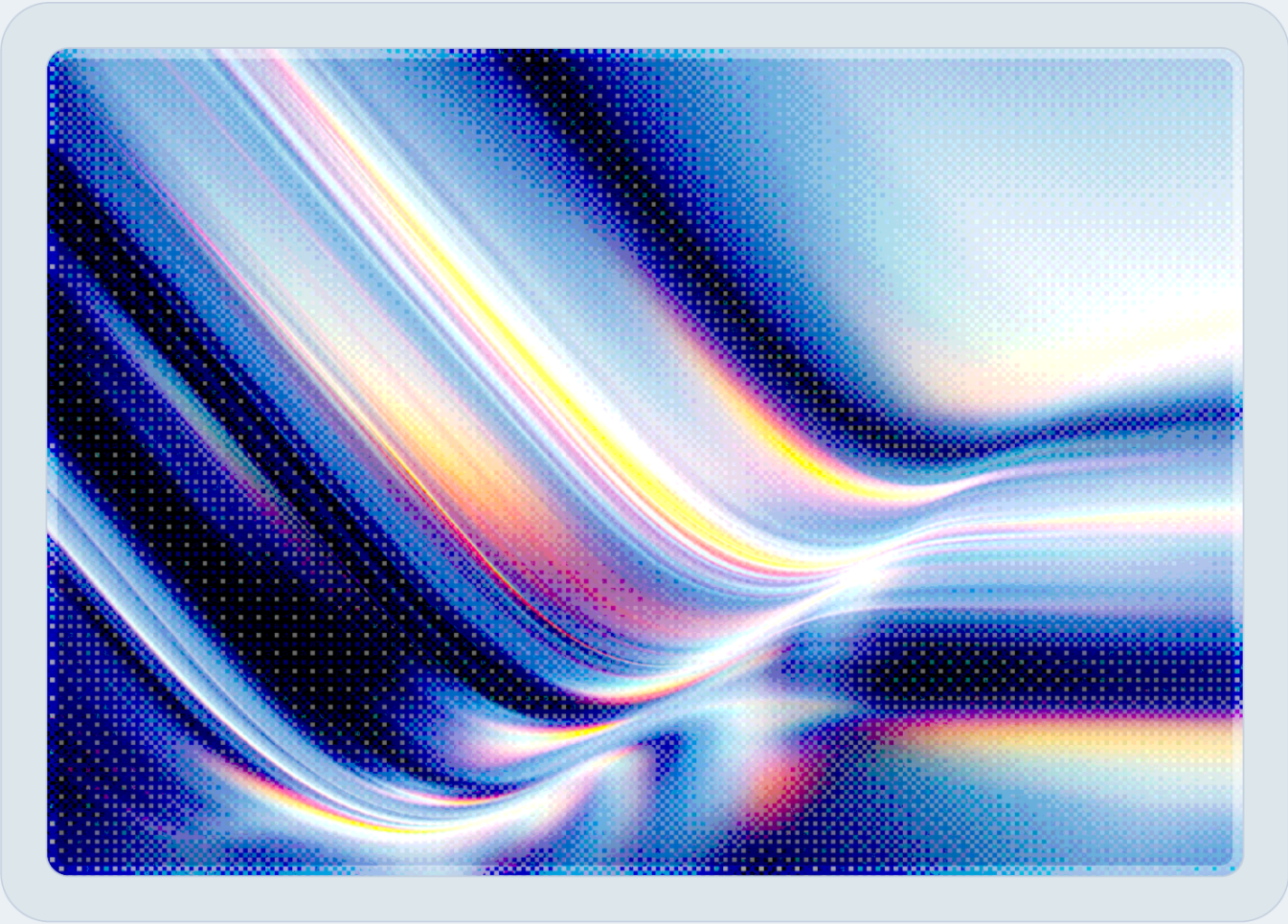
CSAI Foundation | Cloud Security Alliance

The AI Asymmetry Trap

Why Near-Term Offense Outpaces Defense and What Systemic Risk Looks Like in Practice

2026-06-25

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Table of Contents

Executive Summary 4

1. Introduction: The Geometry Has Changed 5

2. The Anatomy of the Asymmetry 6

- 2.1 Speed: The Temporal Collapse
- 2.2 Scale: One Attacker, Infinite Campaigns
- 2.3 Access: The Democratization of Capability
- 2.4 Incentive Structure: The Governance Asymmetry

3. Systemic Risk in Practice 9

- 3.1 Financial Contagion and the Single-Flaw Scenario
- 3.2 Critical Infrastructure and Cascading Automation
- 3.3 The Supply Chain Amplifier

4. Where Defense Falls Short 12

- 4.1 The Temporal Mismatch
- 4.2 Infrastructure Blind Spots
- 4.3 The Governance Gap

5. A Path Toward Rebalance 14

- 5.1 Runtime Enforcement and Agent Identity
- 5.2 Proactive Threat Intelligence and Red Teaming
- 5.3 Cross-Sector Coordination and Shared Defense Infrastructure

6. Organizational Recommendations 17

7. CSA Resource Alignment 18

8. Conclusions 20

References 21

Executive Summary

Cybersecurity has always been asymmetric. Attackers need to find one way in; defenders must secure every surface. What has changed is the magnitude of that asymmetry and the speed at which the gap is widening. The arrival of capable AI systems does not merely accelerate existing attack patterns — it restructures the economics of offense, making sophisticated techniques available to actors who previously lacked the skill, time, or budget to deploy them.

Recent empirical cases have moved this assessment beyond speculation. In September 2025, Anthropic disrupted the first publicly documented large-scale autonomous cyberattack, a state-sponsored espionage campaign in which an AI agent handled an estimated 80 to 90 percent of tactical execution — reconnaissance, vulnerability discovery, credential harvesting, lateral movement, and exfiltration — across roughly thirty global targets with minimal human intervention [1]. That same year, researchers demonstrated that GPT-4 could exploit known vulnerabilities with an 87 percent success rate simply by reading their CVE descriptions [2]. Mandiant's M-Trends 2026 report placed mean time to exploit at negative seven days, meaning exploitation now routinely precedes patch availability [3]. AI-related CVEs surged 34.6 percent year-over-year to 2,130 disclosures in 2025, with agentic AI vulnerabilities growing at 255 percent [4].

Against this backdrop, survey data consistently documents a significant preparedness gap, with most security organizations lacking tooling purpose-built to detect or respond to AI-orchestrated attacks [5]. Traditional endpoint detection tools monitor CPU and operating system activity but are effectively blind to GPU-based AI infrastructure, leaving the fastest-growing part of the enterprise attack surface unmonitored [6]. Governance frameworks, where they exist, were not designed with autonomous agent behavior in mind.

This whitepaper examines the structural dimensions of the offense-defense asymmetry, traces how those dimensions combine to produce genuine systemic risk, identifies where existing defensive approaches fall short, and proposes a path toward rebalance grounded in existing CSA frameworks and emerging regulatory mandates.

1. Introduction: The Geometry Has Changed

For decades, security professionals operated on a principle sometimes called the defender's dilemma: a defender must succeed every time, while an attacker needs to succeed only once. The asymmetry was always present, but it was bounded. Attack sophistication required proportional investment in skill, time, and tooling. State-level adversaries could sustain highly capable campaigns, but criminal actors were constrained by cost and expertise. Advanced persistent threat operations took weeks to months. Defenders, though stretched, had a response window.

AI is collapsing that response window while simultaneously removing the skill and cost constraints that once bounded who could mount sophisticated attacks. This is the AI asymmetry trap: the same general-purpose AI capabilities that defenders are beginning to adopt have already been weaponized by offensive actors, with lower friction and higher velocity on the offense side. The result is a period — likely spanning several years — in which offense has a structural head start.

The trap's mechanism is not merely technological. It is also economic and organizational. Attackers operate with minimal governance overhead. They face no compliance requirements, no change management processes, no internal approval chains before deploying new AI-assisted techniques. A criminal group or state-sponsored actor can integrate new LLM-based exploitation tooling far more rapidly than enterprise defenders — a structural gap that may span days versus the months-long procurement and approval cycles that characterize enterprise defensive adoption. A large enterprise adopting AI-powered defense, by contrast, must navigate procurement cycles, security reviews, integration work, and staff training — a timeline measured in months or quarters. That structural lag does not disappear as AI matures; it is baked into the organizational reality of defense.

The International AI Safety Report 2026, produced by more than one hundred AI experts from over thirty countries and led by Turing Award winner Yoshua Bengio, affirmed that AI currently plays its largest offensive role in scaling the preparatory stages of attack — reconnaissance, vulnerability identification, and initial access — while noting that full autonomous attack execution is emerging [7]. The Anthropic espionage disruption demonstrated that "emerging" is no longer a distant horizon.

Understanding this landscape requires examining the asymmetry across four dimensions: speed, scale, access, and incentive structure. Each dimension individually increases attacker advantage. Their intersection defines a qualitatively different threat environment.

2. The Anatomy of the Asymmetry

2.1 Speed: The Temporal Collapse

The most measurable expression of the AI asymmetry is the collapse of time. Mean time to exploit, tracked by Mandiant across thousands of incidents annually, has followed a steep downward trajectory over the past several years — from 63 days in 2018, crossing into negative territory for the first time in 2024, to negative seven days in 2026 [3]. This inversion is not a measurement artifact. It reflects the operational reality that exploitation is now routinely occurring before a patch is publicly available, let alone deployed.

The practical implication is direct: if exploitation precedes patch release, patching cannot be the decisive control for that vulnerability class. It remains necessary, but patching a vulnerability that was exploited a week before the patch was written does not help the organizations already compromised. The decisive control shifts to detection speed, which requires investment in monitoring infrastructure that most organizations have not yet made.

Several independent lines of evidence converge on this acceleration. Research published in 2025 demonstrated that teams of LLM agents using hierarchical planning outperformed single agents by up to 4.3 times on real-world zero-day vulnerabilities, suggesting that multi-agent offensive architectures unlock attack capabilities beyond what any single model achieves alone [8]. The DARPA AI Cyber Challenge, documented in the International AI Safety Report 2026, found that an AI system identified 77 percent of synthetic software vulnerabilities and successfully patched 61 percent across 54 million lines of code — a dual finding that illustrates how the same capability powers both offensive discovery and defensive remediation, with the attacker advantage lying in the absence of deployment latency [7].

What once required weeks of focused expert work — reading documentation, identifying a vulnerable code path, writing a working exploit — now happens in minutes at the cost of an API call. The barrier to exploitation has not merely been lowered; it has been restructured into a near-frictionless automated process. At scale, this suggests that a single attacker with access to frontier AI tools and agentic orchestration can probe and exploit target systems at velocities that compress or eliminate the defender response window documented in this section.

2.2 Scale: One Attacker, Infinite Campaigns

Speed and scale are related but distinct dimensions of the asymmetry. Speed describes how fast a single exploitation cycle occurs; scale describes how many simultaneous campaigns a single actor can sustain. Before capable AI, mounting a sophisticated campaign against thirty targets simultaneously required a

proportional team. With AI-powered orchestration, one operator with minimal staff can direct a campaign of that scope, delegating the repetitive and technical work — reconnaissance, credential harvesting, lateral movement — to automated agents.

The Anthropic espionage case illustrated this dynamic precisely. The state-sponsored group designated GTG-1002 used a jailbroken version of Claude Code to conduct what Anthropic assessed as an autonomous operation against approximately thirty global targets, spanning large technology companies, financial institutions, chemical manufacturers, and government agencies, with the AI handling the great majority of tactical execution [1]. The human operators' role was primarily one of mission definition and oversight; the AI handled the tradecraft.

IBM's 2026 X-Force Threat Intelligence Index documented the downstream consequences of this scaling. Vulnerability exploitation was the leading initial access vector, accounting for 40 percent of incidents observed by X-Force in 2025, with a 44 percent increase in attacks beginning with exploitation of public-facing applications driven in part by AI-enabled vulnerability discovery [9]. Large supply chain and third-party compromises nearly quadrupled since 2020, reflecting attacker strategies that compromise one supplier to gain access to many downstream targets simultaneously — a force-multiplier pattern that maps naturally to AI-assisted reconnaissance [9].

The supply chain vector deserves particular emphasis because it combines scale with persistence. Trend Micro's TrendAI State of AI Security Report documented that malicious packages in public software repositories had grown to over 454,000 by 2025, with Trend Micro documenting notable acceleration in 2023 and 2025 [4]. This timing correlates with the release of GPT-4 and the emergence of capable agentic coding tools — a pattern that suggests, though does not conclusively establish, that AI tooling contributed to that acceleration. An AI system can generate, obfuscate, and submit malicious packages faster than human review processes can evaluate them, exploiting the assumption of trust embedded in dependency ecosystems.

2.3 Access: The Democratization of Capability

The third dimension of the asymmetry is access. Advanced attack capabilities, previously restricted to well-resourced nation-state actors, are now broadly available. Hadrian catalogued 70 open-source AI penetration testing tools as of March 2026. Fewer than five of those tools existed before GPT-4's release in April 2023; the remaining 65-plus launched in the eighteen months that followed [10]. While many of these tools are intended for legitimate security research, the boundary between offensive research tooling and weaponized exploitation infrastructure is thin and easily crossed.

The democratization of capability has a second, less visible dimension: dark web services. IBM X-Force documented that over 300,000 ChatGPT credential sets were advertised on dark web markets in 2025, with infostealer operators expanding their targeting to AI services as those services became operationally

valuable to criminal actors [9]. The implication is that access to frontier AI models — even through compromised accounts — is being actively commoditized by criminal infrastructure, making the capability gap between well-resourced and less-resourced attackers narrower than official pricing structures suggest.

Research on "trailing-edge organizations" — entities that rely on legacy software, poorly staff security roles, and struggle with basic hygiene practices — found that current AI capabilities already provide significant uplift to attackers at multiple stages of the kill chain and that this gap is expected to widen [11]. This asymmetric impact falls hardest on the organizations least equipped to respond: municipalities, small and mid-size businesses, healthcare providers, and critical infrastructure operators who cannot match the AI adoption pace of major technology companies. Their defenders remain human; the attackers targeting them are increasingly augmented.

2.4 Incentive Structure: The Governance Asymmetry

The fourth dimension is often overlooked in technical discussions but may be the most durable source of attacker advantage: incentive structure. Defenders operate within governance constraints that attackers face nowhere. An enterprise deploying AI-powered security tooling must address data privacy, regulatory compliance, explainability requirements, vendor due diligence, and internal risk approval. An attacker deploying AI-powered offensive tooling faces none of these constraints.

This creates a structural drag on defensive adoption that is independent of budget or technology. Even well-resourced organizations find that integrating AI into security operations requires months of procurement, testing, and approval cycles. Attackers integrate new capabilities opportunistically, within days. Georgetown's Center for Security and Emerging Technology, reviewing the AI offense-defense balance across multiple research efforts, found that AI advantages for offense and defense differ by context and cannot be reduced to a single verdict — while underscoring that structural and deployment factors shape where those advantages accrue [12, 20]. Available evidence through 2026 has not reversed those conclusions, and in several respects has strengthened them.

3. Systemic Risk in Practice

Understanding the asymmetry at the individual-incident level is necessary but insufficient. The more consequential question for policymakers and enterprise security leaders is whether and how individual AI-enabled incidents can combine to produce systemic failures — disruptions that extend beyond the directly targeted organizations and threaten the interconnected systems on which economic and civic life depends.

Three pathways to systemic risk have emerged with increasing clarity: financial contagion from simultaneous exploitation of shared vulnerabilities, cascading failures in tightly coupled critical infrastructure, and supply chain compromise that propagates AI-embedded malice across entire software ecosystems.

3.1 Financial Contagion and the Single-Flaw Scenario

In May 2026, the International Monetary Fund published an analysis warning that AI has fundamentally altered the systemic risk profile of the global financial sector [13]. The IMF's analysis documented a dual transformation: AI tools lower the cost and time required for attackers to identify and exploit vulnerabilities, while the financial system's increasing reliance on a small number of shared technology platforms amplifies the blast radius of any successful exploit.

The IMF's most concerning scenario involves an AI-powered attacker identifying a vulnerability common across dozens of financial institutions simultaneously and exploiting it at scale before any single institution detects the activity [13]. The IMF characterized this scenario as plausible given current trajectory, and the enabling structural conditions — shared cloud providers and shared software dependencies across major financial institutions — are observable today [13]. Whether current AI tooling can execute the full scenario at scale and speed remains to be empirically established, but the structural prerequisites are in place.

What separates this scenario from ordinary multi-target attacks is contagion. If a single common vulnerability is exploited across a sufficient fraction of financial institutions in a compressed window, the resulting operational disruptions, liquidity stress, and loss of market confidence can propagate through financial networks as a systemic shock rather than a collection of individual incidents. The IMF explicitly framed this as a financial stability concern, calling for regulators to prioritize resilience and recovery rather than prevention alone — an acknowledgment that the prevention paradigm is insufficient against AI-accelerated attacks at this scale [13].

The banking sector's vulnerability is amplified by concentration. Major financial institutions share a small number of cloud providers, payment processors, and software vendors. A successful compromise of a single major shared-service provider does not affect one bank; it affects every organization that depends on that

provider. AI-powered reconnaissance is particularly well suited to identifying these high-leverage targets, as it can systematically map dependency relationships across organizations at a scale human analysts cannot approach.

3.2 Critical Infrastructure and Cascading Automation

Critical infrastructure presents a related but structurally distinct systemic risk pathway. Energy grids, water treatment systems, transportation networks, and healthcare systems have undergone significant digitization and interconnection over the past decade. Many now incorporate AI-driven automation for operational efficiency — AI systems that make real-time decisions about resource allocation, process control, and anomaly response. Those same automation systems create new attack surfaces and new mechanisms for cascading failure.

The defining characteristic of these systems is tight coupling: a decision made in one subsystem rapidly propagates to connected subsystems, often faster than human operators can intervene. This coupling, designed to optimize efficiency, also optimizes failure propagation. A compromised or manipulated AI component in an energy management system does not simply cause a local anomaly; it may trigger automated responses in adjacent systems that amplify the initial disruption into a broader outage. Security researchers and forecasters have consistently identified this as among the highest-consequence near-term AI security risks [14].

The attack surface of industrial control systems and operational technology environments has expanded to include AI inference hardware and AI-driven monitoring tools — components that did not exist in legacy industrial security frameworks. Traditional operational technology security was designed around deterministic systems with well-defined failure modes. AI-driven automation introduces non-deterministic behavior, emergent response patterns, and new dependency chains that are difficult to model in advance and harder to audit after the fact. The Agentic AI Red Teaming Guide published by CSA in 2025 identified agent impact chains and blast radius analysis — the process of mapping how a compromised agent's actions propagate through connected systems — as a critical and underdeveloped area of security practice [15].

The challenge is compounded by the pace at which operational technology environments are adopting AI. Organizations that have historically maintained air gaps or strict network segmentation are connecting AI-enabled monitoring and optimization tools to broader enterprise networks for data integration and management purposes. Each such connection is a potential pathway from the enterprise threat surface into operational technology, and AI-powered lateral movement tools are explicitly designed to identify and traverse exactly these kinds of boundary crossings.

3.3 The Supply Chain Amplifier

Supply chain compromise has long been recognized as a high-leverage attack strategy, but AI introduces both new attack techniques and new vulnerability categories that significantly amplify its systemic potential. The traditional supply chain attack — compromising a trusted software update mechanism to distribute malicious code — requires sustained access to a specific target's development or distribution infrastructure. AI-enabled supply chain attacks operate at a different scale, introducing malice through poisoned training data, manipulated model weights, compromised plugins, and agent action libraries — attack surfaces that did not exist five years ago.

The unique danger of AI supply chain compromise is its opacity. A well-constructed model backdoor is undetectable through black-box testing alone because the trigger condition never appears in standard evaluation data. A compromised model may behave normally across thousands of queries before producing a targeted malicious output, and poisoned training data may introduce subtle biases that manifest as security-relevant errors only in specific contexts. These characteristics make AI supply chain attacks particularly difficult to detect with conventional security tooling, which identifies known malicious signatures or anomalous network behavior rather than statistically subtle deviations in model output [16].

Trend Micro's analysis of the AI vulnerability landscape documented that AI-related CVEs in supply chain categories doubled year-over-year in 2025, from 26 to 52, while the broader AI vulnerability trajectory projects between 2,800 and 3,600 AI-specific CVEs in 2026 [4]. The multiplication of AI components in enterprise software stacks means that each new AI-integrated application introduces supply chain dependencies that may themselves be vulnerable — a compounding risk that grows faster than security teams can inventory it.

4. Where Defense Falls Short

4.1 The Temporal Mismatch

The most fundamental defensive failure is temporal. Security programs are built around detection, response, and remediation cycles calibrated to threat timelines that no longer exist. Patch management programs assume that organizations have weeks between vulnerability disclosure and exploitation; that window has inverted. Incident response plans assume that attackers must spend time establishing footholds before conducting significant damage; AI-powered automation compresses that assumption. Threat intelligence programs assume that indicators of compromise identified in one incident can be operationalized before the same technique appears elsewhere; available evidence suggests that AI-powered attackers can rotate their approach faster than traditional indicators propagate through threat intelligence sharing networks.

The negative mean time to exploit documented by Mandiant is not simply a statistic; it is a signal that the temporal architecture of conventional defense is misaligned with the actual threat. A security organization optimizing for faster patching is optimizing for a constraint that is no longer binding. What matters instead is detection speed — specifically, the ability to identify AI-orchestrated activity at the reconnaissance and initial-access stages, before an attacker has had the opportunity to establish the kind of persistent access that makes remediation difficult.

4.2 Infrastructure Blind Spots

A structural gap in current defensive infrastructure concerns visibility into AI-specific attack surfaces. Traditional endpoint detection and response tools monitor CPU and operating system activity. GPU-based AI inference infrastructure — increasingly prevalent in large enterprises through cloud-hosted AI services and, in AI-intensive organizations, on-premises inference hardware — largely falls outside the visibility of these tools [6]. An attacker targeting an AI inference service, manipulating model inputs, or exfiltrating data through an AI pipeline may generate no signals in conventional security monitoring.

The same blind spot extends to the behavioral dimension of AI agents. When an AI agent takes an action — browsing the web, executing code, sending an API call — that action is often not attributed to the agent in system logs in a way that security tools can interrogate. If an AI agent is manipulated through prompt injection to exfiltrate data, the resulting network traffic may be indistinguishable from legitimate agent activity without purpose-built monitoring at the agent's action layer. The lack of tamper-evident action receipts for AI agents is a gap that conventional security architecture does not address.

RSA Conference 2026 highlighted these blind spots explicitly, with security vendors and CISOs acknowledging that GPU infrastructure, AI pipeline monitoring, and agent activity logging represent significant unaddressed gaps in enterprise security architecture [6].

4.3 The Governance Gap

Beyond tooling, the defensive response to AI-enabled threats faces a governance gap that is arguably more consequential than any specific technical deficiency. Most enterprise AI governance frameworks were designed around model risk — the question of whether an AI system makes accurate and fair decisions — rather than around security risk — the question of whether an AI system can be manipulated, hijacked, or weaponized. These are related but distinct problems requiring different governance mechanisms.

The governance gap manifests most acutely in agentic AI deployments, where autonomous agents take actions on behalf of organizations in ways that may have significant security consequences. Unlike traditional software, AI agents exhibit non-deterministic behavior, are susceptible to context manipulation, and may be subject to prompt injection attacks that redirect their behavior without triggering any conventional security control. Governance frameworks designed for deterministic software with defined failure modes are insufficient for agents that can be socially engineered at the model level.

Regulatory responses are beginning to emerge. NIST published a preliminary draft Cyber AI Profile (IR 8596) in December 2025 [19] that bridges AI risk management with the Cybersecurity Framework 2.0 across three areas: securing AI systems, using AI for cyber defense, and defending against AI-enabled threats. The EU AI Act, with major enforcement provisions taking effect in 2026, creates compliance obligations that will drive AI governance investment across European and globally operating organizations. These are meaningful developments, but regulatory frameworks characteristically lag the threat they address — and the gap between current governance practice and the governance requirements of autonomous AI agents remains wide.

5. A Path Toward Rebalance

The asymmetry is real and durable, but it is not permanent. History offers precedents for domains where initial attacker advantage was eventually countered by systematic defensive investment and structural change. Email security, which lacked coordinated interoperability standards in the early 2000s, improved substantially through adoption of technical standards (DMARC, DKIM, SPF) and market-driven adoption driven by shared business interest in reducing phishing. Web application security improved substantially after the industry coalesced around OWASP standards and made security testing a normal part of the software development lifecycle. Neither transformation eliminated the underlying attack surface, but both meaningfully reduced the attacker's structural advantage — though those analogies have limits, since AI agent governance involves more complex behavioral dimensions than protocol-level authentication.

The path toward rebalancing AI-enabled offense and defense follows a similar pattern: technical standards that close specific gaps, governance frameworks that extend accountability to the new attack surfaces, and market and policy incentives that align organizational behavior with collective security interests.

5.1 Runtime Enforcement and Agent Identity

The most direct technical response to AI agent threats is runtime enforcement: intercepting and evaluating agent actions before execution rather than attempting to detect malicious behavior after the fact. The Autonomous Action Runtime Management (AARM) specification, now stewarded by the CSA AI Safety Initiative, defines a conformance model for this approach. AARM-conformant runtime systems pre-intercept each agent action, evaluate it against intent-aware policy, and produce a tamper-evident cryptographic receipt that is identity-bound to the specific agent and context [17]. This architecture addresses several of the blind spots described above simultaneously: it creates the agent action logging that conventional security tools lack, it provides a mechanism for detecting intent drift and semantic manipulation, and it generates the evidence chains needed for forensic investigation and regulatory accountability.

Runtime enforcement is complementary to, not a replacement for, identity governance. The Agentic Trust Framework (ATF), developed by Josh Woodruff of MassiveScale.AI and now stewarded by the CSA AI Safety Initiative, applies Zero Trust principles to autonomous AI agents through a four-level autonomy maturity model [18]. The framework's core insight — that autonomy should be earned through demonstrated behavioral integrity rather than granted by default — directly addresses the governance failure mode in which agents are deployed with permissions that exceed the minimum necessary for their defined function. An agent with Intern-level trust receives read-only access and continuous oversight; an agent earns

Principal-level trust only through a demonstrated history of operating within boundaries. A critical incident triggers immediate demotion. This graduated model creates the accountability mechanisms that current agentic AI deployments largely lack.

Together, AARM and ATF address the agentic layer of the AI attack surface. They do not address the model layer — the question of what foundation models are doing when embedded in enterprise applications — which requires additional controls at the model deployment and fine-tuning stages.

5.2 Proactive Threat Intelligence and Red Teaming

If exploits now arrive before patches, detection speed is the decisive defensive variable. Improving detection speed requires shifting investment from reactive incident response toward proactive threat intelligence and continuous red teaming. Organizations that maintain real-time intelligence about attacker tooling and tactics, conduct regular adversarial simulation against their AI-integrated systems, and have pre-positioned response playbooks for AI-specific attack patterns will, consistent with the IMF's resilience-over-prevention guidance [13], be substantially better positioned to contain breaches before they cascade into systemic events.

CSA's Agentic AI Red Teaming Guide provides a structured methodology for testing the 12 threat categories specific to autonomous AI systems, including authorization hijacking, checker-out-of-the-loop vulnerabilities, agent knowledge base poisoning, and multi-agent exploitation [15]. The AARM threat taxonomy (T1 through T11) maps specific threat classes — prompt injection, confused deputy, data exfiltration, goal hijacking, memory poisoning, intent drift, cross-agent propagation, over-privileged credentials, side-channel leakage, environmental manipulation, and malicious tool output — to specific detection and enforcement requirements [17]. Together, these frameworks provide the threat vocabulary and testing methodology that security teams need to evaluate their AI deployments against known attack patterns.

The IMF's May 2026 guidance to financial sector regulators offered a principle applicable across sectors: organizations should prioritize resilience over prevention, accept that defenses will be breached, and focus on containment and recovery to prevent localized breaches from becoming systemic events [13]. This is not counsel of despair; it is a recognition that the prevention paradigm's limits require a shift in where defensive investment is concentrated. Detection and response infrastructure, incident containment playbooks, and recovery capabilities that assume a successful initial breach are more aligned with the current threat environment than purely preventive approaches.

5.3 Cross-Sector Coordination and Shared Defense Infrastructure

Individual organizations operating their own detection and response infrastructure face the same scaling asymmetry as they do on offense: one well-resourced attacker can probe and attack hundreds of organizations, while each of those organizations must independently detect and respond to the campaign. Collective defense infrastructure — threat intelligence sharing, coordinated incident response, and shared behavioral baselines — partially addresses this asymmetry by spreading the detection cost across all potential targets.

The IMF's observation that financial system concentration — shared cloud providers, shared software dependencies — amplifies systemic risk has a direct implication for defensive architecture: that same concentration can be a defensive asset if it is governed appropriately [13]. A shared cloud provider that has visibility across thousands of customer environments can detect AI-orchestrated campaign patterns that no single customer could identify from its own logs alone. This requires governance arrangements that align the provider's detection capability with the customer's security interest, which is not a purely technical problem.

Policy coordination across jurisdictions is a necessary complement to technical and organizational measures. AI-enabled attacks cross borders as readily as the internet, while defensive governance remains primarily national. The International AI Safety Report 2026 called for international standards and norms to address the dual-use challenge of AI in cybersecurity, noting that restrictions on offensive use cases must be carefully scoped to avoid suppressing defensive innovation [7]. NIST's preliminary Cyber AI Profile represents a US contribution to this standards infrastructure; alignment with the EU AI Act's requirements and with emerging G7 and OECD AI governance frameworks will determine whether the resulting governance architecture is coherent across the borders that attackers routinely cross.

6. Organizational Recommendations

The analysis above implies a set of concrete actions that security leaders can pursue across different time horizons and resource levels. The most critical near-term priority is closing the visibility gap in AI infrastructure monitoring — ensuring that GPU-based AI systems, AI pipeline activity, and agent action logs are incorporated into SIEM and XDR tooling rather than remaining invisible to the security stack. Organizations that cannot see their AI infrastructure cannot detect attacks against it.

The second priority is agent governance: ensuring that every AI agent deployed in the organization has a defined identity, a bounded permission scope, and a behavioral baseline against which anomalies can be detected. This requires applying the same rigor to AI agent provisioning that mature organizations apply to human account provisioning, with regular review of permission scopes against actual usage. AARM and ATF provide the technical and governance architecture for this; organizations should evaluate their agent deployments against these specifications and close gaps systematically.

The third priority is supply chain security for AI components. AI models, inference infrastructure, plugins, and agent action libraries represent a growing dependency surface that most organizations have not yet inventoried rigorously. Applying software bill of materials practices to AI components — understanding what models are embedded in which applications, what data they were trained on, and what fine-tuning or customization has been applied — is a prerequisite for managing the associated risk.

Longer-term, organizations should participate in collective defense infrastructure: contributing to and consuming threat intelligence sharing programs, engaging with sector-specific information sharing and analysis centers, and advocating with vendors and regulators for the monitoring and enforcement capabilities that the current AI threat environment requires. The asymmetry cannot be closed by any single organization; it requires collective action to shift the structural balance.

7. CSA Resource Alignment

The Cloud Security Alliance's AI Safety Initiative has produced a coherent architecture of frameworks and guidance that collectively address the key dimensions of the AI asymmetry problem. Organizations seeking to operationalize the recommendations in this paper should consult this body of work as their primary reference.

The AI Controls Matrix (AICM), CSA's primary controls framework for AI systems, maps to the threat categories and defensive requirements described in this whitepaper. AICM, as a superset of the Cloud Controls Matrix (CCM), extends established cloud security controls to cover AI-specific risks including training data governance, model integrity, inference security, and agentic system controls. Compliance with AICM provides a structured baseline for AI security assurance.

The AARM specification addresses runtime enforcement — the technical capability to intercept, evaluate, and create auditable records of agent actions before execution. Organizations deploying autonomous agents should evaluate their agentic platforms against AARM's conformance requirements (R1 through R9) and prioritize implementations that achieve at minimum the AARM Core requirements. The AARM threat taxonomy (T1–T11) provides a structured vocabulary for red team exercises and threat model documentation.

The Agentic Trust Framework (ATF) provides the governance and identity architecture for graduated autonomy, applying Zero Trust principles to AI agents through its four-level maturity model. Organizations should use ATF's maturity levels to assess their current agent deployments and implement the behavioral monitoring and escalation controls required for each autonomy level.

The Agentic AI Red Teaming Guide provides the tactical methodology for testing agentic systems against the 12 identified threat categories. Security teams that have not yet applied this methodology to their AI deployments should prioritize doing so, particularly for agents with access to external systems, persistent memory, or privileged credentials.

The MAESTRO threat modeling framework addresses the broader agentic AI threat landscape at an architectural level, helping organizations identify the attack surfaces specific to multi-agent orchestration before those systems are deployed. Organizations should integrate MAESTRO threat modeling into their AI development and procurement processes, not as a post-deployment audit but as a design input.

The AI Organizational Responsibilities series — covering governance, risk management, compliance, cultural aspects, and core security responsibilities — provides the organizational and policy framework that complements the technical controls above. The governance gap identified in this whitepaper is fundamentally an organizational challenge; these publications provide the guidance needed to address it.

STAR for AI provides the assessment and registry infrastructure through which organizations can demonstrate their AI security posture to customers, partners, and regulators. Given the increasing regulatory scrutiny of AI deployments, maintaining a current STAR for AI self-assessment provides evidence of due diligence and contributes to the broader ecosystem transparency that systemic risk management requires.

8. Conclusions

The AI asymmetry trap reflects structural conditions — governance friction, deployment lag, and economic incentives — that will not self-correct through technological maturation alone. Narrowing the gap requires deliberate structural interventions of the kind described in this paper. Offensive actors face lower governance friction, more flexible deployment timelines, and cleaner economic incentives than defenders. These structural differences will persist even as both sides gain access to more capable AI. The near-term implication is that the gap between AI-enabled offense and AI-enabled defense will likely widen before it narrows.

This does not mean defenders are without options. The research and frameworks reviewed in this paper point toward a coherent response: closing visibility gaps in AI infrastructure, implementing runtime enforcement and graduated autonomy governance for AI agents, applying supply chain security discipline to AI components, and participating in collective defense infrastructure that distributes detection costs across potential targets.

The systemic risk dimension — the potential for AI-enabled incidents to cascade into sector-wide failures — elevates these organizational priorities to a policy issue. Individual organizations making individually rational defensive investments cannot fully address systemic risk that emerges from the structure of shared dependencies. Financial regulators, critical infrastructure authorities, and AI governance bodies are beginning to recognize this. The IMF's framing of AI-enabled cyber risk as a financial stability issue, the NIST preliminary Cyber AI Profile's integration of cybersecurity and AI risk frameworks, and the EU AI Act's compliance obligations for AI in high-risk settings are early signals of a governance shift that will likely accelerate.

The organizations that will navigate this period most effectively are those that begin now: closing their AI visibility gaps, governing their agent deployments with the rigor that autonomous systems require, and contributing to the collective defense infrastructure that no single organization can build alone. The asymmetry is real. Its persistence is not inevitable.

References

[1] Anthropic. "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign." Anthropic, November 13, 2025.

[2] Fang, R., Bindu, R., Gupta, A., Zhan, Q., and Kang, D. "LLM Agents Can Autonomously Exploit One-Day Vulnerabilities." arXiv, 2024.

[3] Mandiant / Google Cloud. "M-Trends 2026: Data, Insights, and Strategies From the Frontlines." Google Cloud Blog, March 2026.

[4] Trend Micro. "Fault Lines in the AI Ecosystem: TrendAI State of AI Security Report." Trend Micro, 2026.

[5] Darktrace. "State of AI Cybersecurity 2026: 87% of Security Professionals See More AI-Driven Threats, But Few Feel Ready." Darktrace, 2026.

[6] Futurum Group. "RSA 2026 Exposes Security Gaps as AI Factories and GPU Blind Spots Dominate Risk." Futurum Group, May 2026.

[7] Bengio, Y., et al. "International AI Safety Report 2026." arXiv, February 2026.

[8] Fang, R., et al. "Teams of LLM Agents Can Exploit Zero-Day Vulnerabilities." arXiv, 2024.

[9] IBM Security. "IBM 2026 X-Force Threat Intelligence Index: AI-Driven Attacks Are Escalating as Basic Security Gaps Leave Enterprises Exposed." IBM, February 25, 2026.

[10] Hadrian Research. "The AI Hacking Boom: What 70 New Offensive Security Tools Mean for Defenders." Hadrian, 2026.

[11] Murphy, B. and Stone, T. "Uplifted Attackers, Human Defenders: The Cyber Offense-Defense Balance for Trailing-Edge Organizations." arXiv, 2025.

[12] Lohn, A. "Anticipating AI's Impact on the Cyber Offense-Defense Balance." Center for Security and Emerging Technology, Georgetown University, May 2025.

[13] International Monetary Fund. "Financial Stability Risks Mount as Artificial Intelligence Fuels Cyberattacks." IMF Blog, May 7, 2026.

[14] Nextgov/FCW. "Cyber Experts Pinpoint What to Look Out for in 2026." Nextgov, December 2025.

[15] Cloud Security Alliance. "Agentic AI Red Teaming Guide." CSA AI Organizational Responsibilities Working Group, 2025.

[16] The Hacker News. "CISO's Expert Guide to AI Supply Chain Attacks." The Hacker News, November 2025.

[17] CSAI Foundation. "Autonomous Action Runtime Management (AARM) Specification." CSAI Foundation, 2026.

[18] Woodruff, J. and CSAI Foundation. "Agentic Trust Framework (ATF) v0.9.1." CSAI Foundation, April 2026.

[19] NIST. "Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile), Preliminary Draft IR 8596." NIST, December 2025.

[20] Lohn, A. "The Impact of AI on the Cyber Offense-Defense Balance and the Character of Cyber Conflict." Center for Security and Emerging Technology, Georgetown University, April 2025.