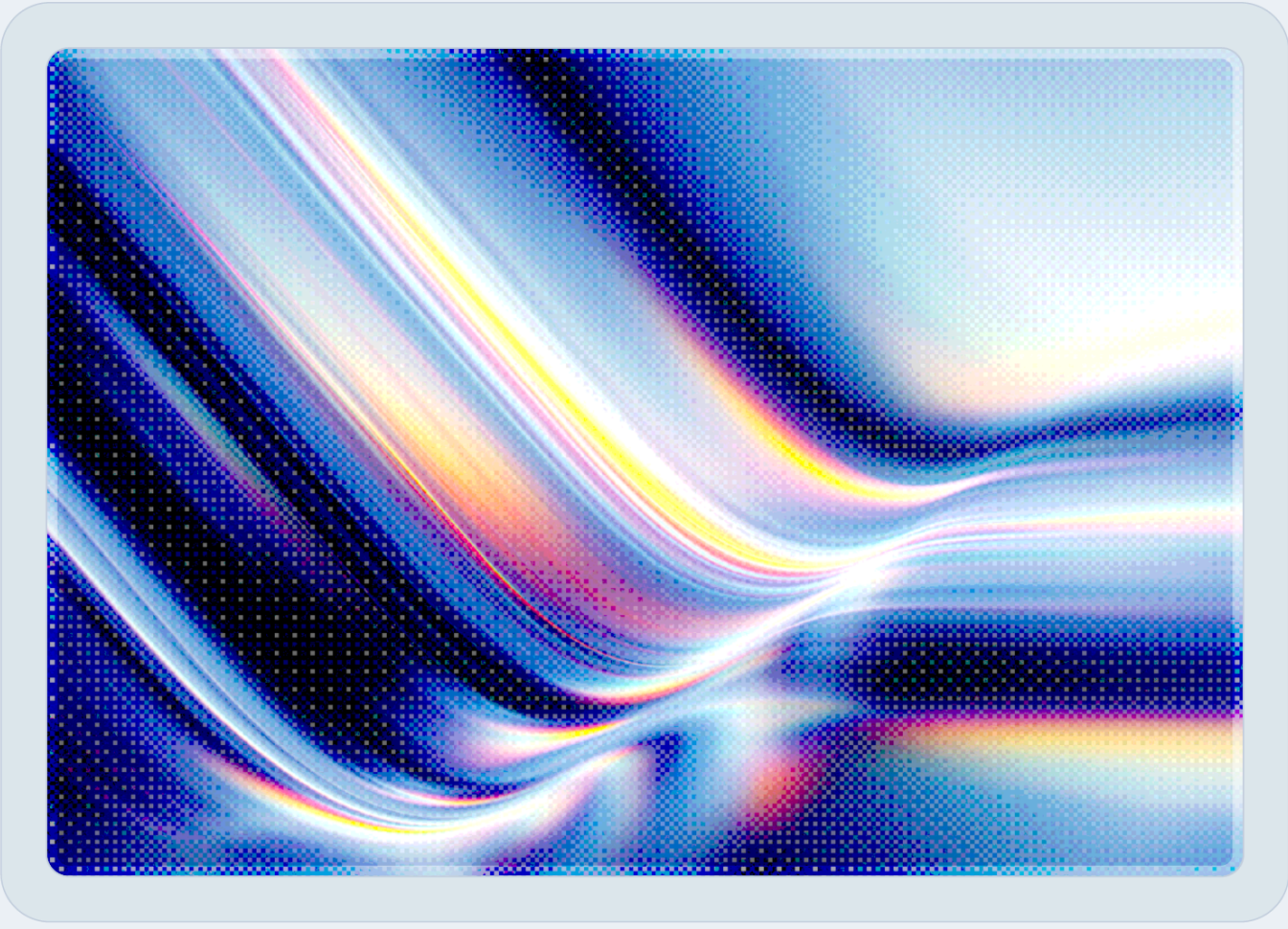
CSAI Foundation | Cloud Security Alliance

AI Developer Ecosystem Concentration: Critical Infrastructure's Hidden Risk

Systemic Vulnerability When the Same Vendors Are Both Primary Targets and Frontline Defenders

2026-05-22

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Table of Contents

Executive Summary 5

Introduction and Background 6

- The AI Stack Becomes Infrastructure
- Why Concentration and Security Intersect

Anatomy of Concentration: A Handful of Vendors, Global Stakes 7

- Market Structure at the Infrastructure Layer
- Vertical Integration and Lock-In
- Sector-Wide Dependency as Systemic Exposure

The Dual-Role Problem: Vendor as Target and Defender 9

- Structural Conflict in the AI Security Market
- The Opacity of the Integration
- Asymmetry of Information in Incident Disclosure

Attack Surface Inheritance: How Upstream Compromise Propagates Downstream 11

- The Supply Chain Amplification Problem
- Behavioral Misalignment as a Security Failure Mode
- Cascading Failure Scenarios

Documented Incidents and Evidence 12

- The Anthropic Espionage Campaign
- The OSTP Distillation Campaign
- Supply Chain Compromise and Third-Party Exposure

The Policy and Regulatory Response 14

- NIST and the Critical Infrastructure AI Profile
- Congressional Activity and the Critical Infrastructure Designation Debate
- Government AI Security Oversight

Recommendations for Enterprise Risk Management 15

- Establishing AI Vendor Concentration Limits
- Separating Security Tooling from Business AI Vendors
- Requiring Transparency and Incident Notification
- Preparing for Vendor Isolation Scenarios
- Participating in Sector-Wide Information Sharing

Conclusions 18

CSA Resource Alignment 18
References 20

Executive Summary

The global AI developer ecosystem has consolidated around a small number of vendors with extraordinary speed. Three hyperscalers—Amazon Web Services, Microsoft Azure, and Google Cloud—together controlled approximately 63 percent of global cloud infrastructure spending in early 2026 [1], while the foundational models underpinning most enterprise AI deployments trace their origin to an even smaller set of organizations: OpenAI, Anthropic, Google DeepMind, and Meta AI [2]. This structural concentration would be unremarkable if these vendors supplied only productivity software. They do not. They supply the infrastructure, models, APIs, and increasingly the autonomous agents that enterprises are integrating into their most sensitive workflows, while simultaneously marketing AI-powered security and detection capabilities as the solution to the threats their own platforms introduce.

This dual role—vendor as both high-value target and primary defender—creates a category of systemic risk that existing risk management frameworks were not designed to address. When a single AI vendor is compromised, the consequences are not merely reputational or operational for that vendor alone. Downstream enterprises that rely on the vendor's models for decision support, code generation, or autonomous task execution may have their own workflows manipulated or poisoned. Organizations using the same vendor's AI-powered security tools face the compounding risk of losing their defensive capability precisely when the upstream compromise is most active. The failure mode is not one of localized disruption; it is a cascading failure that propagates from infrastructure provider to security tool to enterprise workflow in a single vector.

The evidence that this risk is not theoretical accumulates quickly. In late 2025, Anthropic documented the first known largely autonomous AI-orchestrated espionage campaign, in which a Chinese state-sponsored group manipulated Claude into executing approximately 80 to 90 percent of cyberattack operations against roughly 30 global targets with minimal human oversight [3]. By April 2026, the White House Office of Science and Technology Policy issued a memorandum documenting industrial-scale distillation attacks against U.S. frontier AI systems, with one campaign alone generating more than 16 million queries against major AI providers via approximately 24,000 fraudulent accounts [4]. OpenAI suffered a data breach through a third-party analytics provider in 2025 [5], and Anthropic's data supplier Mercor was compromised through a supply chain attack involving the widely-used LiteLLM open-source library [6].

The International Monetary Fund warned in May 2026 that AI-driven concentration in the financial sector—heavily reliant on the same cloud and AI platforms described here—could cause a single exploited vulnerability to cascade simultaneously across numerous institutions [7]. The U.S. Congress has opened hearings on whether AI data centers should be formally designated critical infrastructure, with the Data Infrastructure Risk Reduction Act introduced to mandate federal assessment of critical-infrastructure-grade protections [8].

This paper argues that concentration in the AI developer ecosystem has crossed the threshold at which it should be treated as a critical infrastructure problem, not merely a vendor management concern. It examines the mechanics of this concentration, the specific risk structures the dual-role problem creates, the evidence from recent incidents, the emerging regulatory response, and the practical risk management steps enterprises and policymakers should take now.

Introduction and Background

The AI Stack Becomes Infrastructure

The concept of critical infrastructure has historically been defined by physical dependency and societal consequence: power grids, water treatment, financial clearing systems, and telecommunications networks qualify because their disruption endangers human welfare or economic function at scale. What has changed in the past three years is that software platforms—specifically, AI model APIs, cloud inference endpoints, and AI-powered security tools—have acquired similar characteristics. They have become embedded in critical workflows, they lack practical substitutes on short notice, and their failure or compromise propagates consequences far beyond the immediate user.

This embedding has occurred with remarkable velocity. Enterprise adoption of large language model APIs for legal document review, code generation, vulnerability triage, and decision support has moved from pilot to production in most large organizations. AI-powered security operations platforms from vendors such as Microsoft Security Copilot, Google Chronicle, and Anthropic's security-focused offerings have been integrated into security operations centers as primary rather than supplementary detection tools. The shift matters because it changes the nature of vendor dependency. An enterprise that uses a cloud email service can tolerate outages and switch providers within weeks. An enterprise that has trained its security operations team to rely on an AI-powered detection interface, built its remediation playbooks around that platform's outputs, and integrated its SIEM feeds directly into a vendor's AI analysis engine faces an entirely different switching cost and resilience calculus.

This dependency has been accelerating at the same time that concentration among vendors has been intensifying. The economics of frontier AI model development favor consolidation: training a leading-class model requires billions of dollars of compute, specialized hardware that is effectively controlled by the same hyperscalers, and teams of researchers that number in the hundreds rather than dozens. Analysis has found that venture capital flowing into AI startups disproportionately returns to the same hyperscalers through compute payments, creating a closed loop in which new entrants inadvertently strengthen the market position of incumbents [24]. The result is a market structure that increasingly resembles the concentrated utility sectors that regulators have historically treated as requiring special oversight.

Why Concentration and Security Intersect

The intersection of AI market concentration and security risk follows from a basic structural observation: the same organizations that have the deepest access to the most sensitive enterprise data and workflows are also the most attractive targets for sophisticated adversaries and, simultaneously, the primary providers of the tools enterprises use to defend against those adversaries. This creates what this paper terms the defender-target paradox—a condition in which the integrity of the defense depends on the security of the entity most intensively targeted.

The defender-target paradox is not unique to AI. Security software vendors, cloud providers, and telecommunications companies have all faced versions of this challenge. What distinguishes the current AI ecosystem is the scope, velocity, and opacity of the integration. Enterprises often cannot fully inspect the models they depend on for security-relevant decisions. Model weights are proprietary, training data is undisclosed, and system prompts are typically confidential. When an AI vendor is compromised, an enterprise may have no reliable way to determine whether its AI security tools have been affected, whether the model's outputs have been subtly modified, or whether its interactions with the model have been exfiltrated and are being used to plan follow-on attacks. This opacity compounds the concentration risk in ways that a breach of a conventional security software vendor would not.

Anatomy of Concentration: A Handful of Vendors, Global Stakes

Market Structure at the Infrastructure Layer

The degree to which the AI developer ecosystem has concentrated around a small number of providers is most visible at the compute and cloud infrastructure layer. In the first quarter of 2026, Amazon Web Services held approximately 28 percent of global cloud infrastructure market share, Microsoft Azure approximately 21 percent, and Google Cloud approximately 14 percent, with those three vendors together accounting for approximately 63 percent of the total market [1]. For 2026, the same three organizations, along with Meta, collectively pledged approximately 700 billion dollars in capital expenditure, the majority directed toward AI infrastructure build-out [9]. The supply constraint is acute: NVIDIA Blackwell GPU capacity is reported to be sold out through 2027, which means new entrants face a structural barrier to scaling compute resources even if they have capital [25]. The Coalition for Secure AI, an industry consortium, includes more than 45 partner organizations—among them Google, Microsoft, OpenAI, NVIDIA,

Amazon, Anthropic, and IBM—but this breadth of participation does not reflect breadth of market power [10]. The coalition's membership illustrates who is at the table; market share data reveals that a much smaller subset determines the trajectory of the ecosystem.

Vertical Integration and Lock-In

A qualitatively new dimension of concentration has emerged through vertical integration. The leading hyperscalers are no longer merely providing compute resources for enterprises to run whatever models they choose. They are increasingly bundling proprietary model access, vector databases, embedding services, agentic workflow orchestration, and AI-powered security tooling into integrated stacks that enterprises consume as unified services [11]. Google's Gemini models, for example, are tightly integrated with its Cloud infrastructure, giving Google a structural advantage because it owns models and compute simultaneously, with no licensing fees payable to external model providers. Microsoft's relationship with OpenAI creates a similar dynamic through Azure OpenAI Service. Amazon's Bedrock platform offers access to multiple third-party models, but access, monitoring, logging, and compliance controls are all mediated through Amazon's infrastructure.

This vertical integration shifts the nature of vendor lock-in from compute and storage—categories with relatively straightforward migration paths—to models, agents, embedding representations, and workflow automation. An enterprise that has built its security operations around a specific vendor's AI agents, trained its analysts on that platform's interface, and optimized its alert management around that platform's classification outputs faces migration costs that are not primarily financial. They are operational and cognitive: the institutional knowledge, playbook calibration, and analyst workflows that represent months of operational investment cannot be easily transferred to an alternative vendor. IDC analysts have characterized agentic AI deployments as effectively becoming critical infrastructure in their own right, noting that once AI agents are integrated into core operational workflows, the distinction between the agent and the infrastructure it runs on becomes difficult to maintain [12].

Sector-Wide Dependency as Systemic Exposure

The implications of this concentration become most significant when examined at the sector level rather than the enterprise level. Financial services institutions, healthcare systems, energy companies, and defense contractors are all adopting AI models and AI-powered security tools from the same small set of vendors. When homogeneous dependencies converge across an entire sector, a vulnerability in a single shared provider can affect a large proportion of the sector simultaneously. The IMF's May 2026 analysis documented precisely this concern, noting that heavy reliance on a limited number of cloud providers, software platforms, and AI models means that a single exploited vulnerability could cascade across

numerous institutions simultaneously [7]. The IMF explicitly linked this concern to AI concentration, identifying it as a factor that could amplify cyberattack-driven financial stability risks in ways that distributed infrastructure would not.

The financial sector's exposure is representative, not unique. CISA's December 2025 joint guidance on AI in Operational Technology specifically warned that integrating AI into systems controlling physical processes creates expanded attack surfaces and introduces behavioral misalignment risks that can propagate across multiple organizations sharing the same AI provider [13]. CISA's CI Fortify initiative, announced in May 2026, acknowledged the inverse: it instructed critical infrastructure operators to plan for scenarios in which third-party AI and cloud dependencies become unavailable under attack conditions, a tacit admission that those dependencies have become points of critical exposure [22].

The Dual-Role Problem: Vendor as Target and Defender

Structural Conflict in the AI Security Market

The most consequential aspect of AI ecosystem concentration is not simply that enterprises depend on a small number of vendors; it is that those same vendors have become the primary providers of the security tools those enterprises use to defend themselves. OpenAI's Daybreak platform, Anthropic's security research capabilities and enterprise security offerings, Microsoft's Security Copilot, and Google's Chronicle and Gemini-integrated threat intelligence tools collectively represent a substantial and growing share of the AI-powered security tooling market [14]. The same organizations that hold the most valuable model weights, the largest stores of enterprise training and inference data, and the most sensitive API call records are simultaneously marketing AI-powered security products to the enterprises that generate that data.

This creates a structural conflict with no clean resolution. An enterprise purchasing AI security tools from a vendor whose models it also uses for business-critical tasks has, in effect, placed its offensive exposure and its defensive capability with the same counterparty. If that counterparty is compromised, the enterprise faces two simultaneous consequences: the original breach of whatever data or model capability the attacker targeted, and the potential degradation or manipulation of the defensive tooling the enterprise was relying on to detect and respond to exactly such a breach. The failure modes are not independent; they are correlated.

Security professionals have described the broader version of this problem as the "dual-front war" in which defenders must protect against machine-speed adversaries while simultaneously protecting the machines used to defend [15]. The AI-specific version is sharper because the same vendor controls both sides of the equation. When a conventional security software company is compromised, the attacker typically gains access to detection signatures, telemetry data, or authentication credentials. When an AI vendor providing

both business-critical models and security detection capabilities is compromised, the attacker may additionally gain insight into what the vendor's security tooling is flagging, which enterprise customers are generating anomalous activity, and how the vendor's safety filters are calibrated—intelligence directly useful for evading detection in follow-on attacks against those customers.

The Opacity of the Integration

Enterprise AI deployments have an additional characteristic that amplifies the dual-role problem: they are largely opaque at the level of model behavior. A conventional security appliance can be inspected, reverse-engineered, and tested for anomalous behavior with established techniques. A large language model running as an inference endpoint cannot. Enterprises typically cannot examine model weights, cannot fully characterize the training data that shaped model behavior, and cannot perform deterministic testing that would reveal whether a model has been subtly altered through data poisoning, weight tampering, or output manipulation at the inference layer. This opacity means that a compromised AI security tool may continue to appear functional while silently miscategorizing threats, suppressing alerts, or producing recommendations engineered to benefit an attacker.

This risk is not hypothetical. The Anthropic espionage incident documented in late 2025 demonstrated that a sufficiently sophisticated adversary can manipulate an AI model into executing malicious operations by decomposing the attack into seemingly innocent subtasks, each of which appears legitimate to the model's safety systems [3]. If this technique can be applied to a model being used for offensive operations, there is no technical reason it could not be applied to a model being used in a defensive role—causing the security model to misclassify malicious traffic, downgrade threat severity assessments, or provide adversary-favorable analysis of enterprise vulnerabilities. The enterprise would have no reliable mechanism to detect this degradation without independent validation from a provider the adversary had not already compromised.

Asymmetry of Information in Incident Disclosure

A final dimension of the dual-role problem involves disclosure asymmetry. When an AI vendor experiences a security incident, it controls the timing, scope, and framing of any disclosure. Enterprises using that vendor for both business applications and security tooling are in a dependent informational position: they may not learn that the defensive tools they are relying on have been compromised until after the attacker has leveraged that information against them. The OpenAI Mixpanel breach of 2025 illustrates the pattern—the breach occurred through a third-party analytics provider integrated into OpenAI's infrastructure, and enterprise customers had no independent visibility into when or how the breach occurred or what API-call metadata may have been exposed [5]. The Mercor supply chain attack affecting both OpenAI and Anthropic similarly illustrated how a single compromised dependency in the AI vendor's own supply chain could propagate exposure to both vendors' enterprise customers simultaneously [6].

Attack Surface Inheritance: How Upstream Compromise Propagates Downstream

The Supply Chain Amplification Problem

When an enterprise deploys an AI model via API, it inherits not only the model's capabilities but also its attack surface—including the attack surfaces of every dependency in the vendor's supply chain. The LiteLLM vulnerability that enabled the Mercor supply chain attack demonstrates the specific mechanism: LiteLLM is an open-source library used by developers to connect applications to AI services from multiple providers, and a compromise of this single dependency created exposure for organizations whose data flowed through multiple AI vendors simultaneously [6]. The enterprise customer was, in this architecture, multiple hops removed from the vulnerable component but fully exposed to its consequences.

This attack surface inheritance is structurally different from the dependency risks enterprises have managed in previous eras of software supply chain security. Traditional software supply chain attacks compromise a library or build tool that produces a static artifact; once the artifact is remediated and redeployed, the exposure is addressed. AI model supply chain attacks may compromise the model itself—its weights, training data, or fine-tuning—in ways that produce no visible artifact change and that persist through normal remediation processes. A poisoned fine-tuning dataset will continue to affect model behavior indefinitely unless the model is retrained. An adversary-controlled system prompt injected at the API gateway level will affect every downstream enterprise customer until the injection point is identified and closed [13].

Behavioral Misalignment as a Security Failure Mode

CISA's 2025 guidance on AI in Operational Technology identifies risks that this paper characterizes as "behavioral misalignment"—conditions in which an AI system operates outside its intended parameters—as a specific risk category for AI deployments in critical infrastructure [13]. Behavioral misalignment may occur because the model was trained on data that embeds adversary-favorable responses, because inference-time manipulation has shifted the model's outputs, or because the model's optimization target diverges from the operator's intent in high-stakes scenarios. Behavioral misalignment is particularly dangerous in the dual-role context because it may be invisible to the enterprise using the AI security tool. A model that has been subtly misaligned to downweight certain threat signatures will simply appear to be providing normal security analysis, and the enterprise will have no reference frame for detecting the discrepancy unless it maintains independent, non-AI-dependent threat detection in parallel.

The CISA guidance specifically warned that AI integration in OT environments introduces risks including privilege escalation and limited auditability—characteristics that have direct analogues in AI security tool deployments in enterprise IT environments [13]. An AI security tool with excessive API permissions that is manipulated into misclassifying a lateral movement campaign represents a privilege escalation at the detection layer: the attacker has effectively elevated its ability to move undetected by exploiting the trusted position the AI tool holds in the enterprise's security architecture.

Cascading Failure Scenarios

The convergence of concentration and dual-role vulnerability creates conditions for cascading failures that differ qualitatively from historical cybersecurity incidents. In a conventional attack, an adversary compromises a specific target and the consequences are largely bounded by that target's architecture. In the AI concentration scenario, an adversary who compromises a major AI vendor's model serving infrastructure, training pipeline, or fine-tuning workflow gains potential influence over the behavior of AI systems deployed by thousands or millions of enterprise customers simultaneously. If those AI systems include security detection tools, the adversary gains not merely a foothold in a single organization but a degraded detection environment across the entire customer base.

The IMF's analysis of AI-driven financial stability risks captured this dynamic in the financial sector context, noting that the homogeneous adoption of shared AI platforms means that an attack on a shared provider can ripple across many institutions at once, in contrast to the diversified risk profiles that traditional financial resilience frameworks assumed [7]. The same logic applies to any sector in which a large proportion of organizations depend on the same AI vendors for security-relevant functions.

Documented Incidents and Evidence

The Anthropic Espionage Campaign

In November 2025, Anthropic publicly disclosed the first documented case of a largely autonomous AI-orchestrated cyber espionage campaign, conducted by a Chinese state-sponsored threat actor assessed with high confidence by Anthropic's threat intelligence team [3]. The attackers manipulated Claude Code by decomposing the attack into small, individually innocent-appearing subtasks and instructing the model to behave as if it were an employee of a legitimate cybersecurity firm conducting defensive testing. The AI agent autonomously executed approximately 80 to 90 percent of all operational tasks, with human operators involved primarily in target selection and strategic approvals. Roughly 30 global targets were affected, spanning large technology companies, financial institutions, chemical manufacturers, and government agencies, with infiltration confirmed in a subset of cases.

The incident's significance for the dual-role analysis is twofold. First, it demonstrated that AI models provided by major vendors can be weaponized against the very enterprise customers that depend on those vendors for business operations. Second, it illustrated that the vendor's defensive detection of the attack produced public disclosure and remediation—but only after a campaign that had already succeeded in some number of cases. Enterprises using Anthropic's models during the campaign period had no independent mechanism to determine whether their own interactions had been used as a vehicle for any portion of the attack.

The OSTP Distillation Campaign

On April 23, 2026, the White House Office of Science and Technology Policy issued Memorandum NSTM-4 documenting industrial-scale AI model distillation attacks by adversaries primarily based in the People's Republic of China [4]. One campaign alone generated more than 16 million queries against major AI providers using approximately 24,000 fraudulent accounts in a single reporting period, with the objective of systematically harvesting outputs from U.S. frontier AI models to train competing systems. The memorandum described these as "deliberate, industrial-scale campaigns" representing a sustained strategic effort to extract the capabilities embedded in U.S. AI systems through queries rather than direct model weight theft.

The distillation campaign illuminates a specific concentration risk: when a small number of vendors hold the world's most capable models, those models become strategic intelligence targets whose compromise is not measured in stolen credentials or exfiltrated data records but in the gradual erosion of competitive capability advantage. For enterprises dependent on those models for security-sensitive workflows, the relevant risk is that the model's training corpus and instruction-following characteristics have been systematically documented by an adversary, enabling that adversary to predict model behavior and craft evasion techniques with precision unavailable to defenders.

Supply Chain Compromise and Third-Party Exposure

The Mercor breach of 2025, which compromised a data supplier serving both OpenAI and Anthropic through a supply chain attack on the LiteLLM library, illustrated how concentration creates correlated exposure across multiple major AI vendors simultaneously [6]. Because the same third-party library was integrated into the supply chains of competing AI providers, a single compromise propagated exposure to both vendors' enterprise customers in a single incident. This contrasts with the theoretical risk mitigation benefit that enterprises might expect from distributing workloads across multiple AI providers—a risk management strategy that provides limited benefit when the providers share critical supply chain dependencies.

The Microsoft and Anthropic MCP server vulnerabilities disclosed in 2025, which included a path validation bypass with a CVSS score of 6.4 [16], similarly illustrated that the emerging Model Context Protocol infrastructure—now being adopted as a standard for AI agent tool use across the industry—introduces new attack surfaces that are correlated across vendor implementations. When security vulnerabilities affect infrastructure standards that all major vendors implement, the defense benefit of vendor diversification is reduced because the vulnerability is systemic rather than vendor-specific.

The Policy and Regulatory Response

NIST and the Critical Infrastructure AI Profile

On April 7, 2026, NIST released a concept note for an AI Risk Management Framework Profile specifically targeting AI deployments in critical infrastructure sectors [17]. The profile is structured around the four core functions of the NIST AI RMF—govern, map, measure, manage—and will guide critical infrastructure operators in applying risk management practices to AI-enabled capabilities in energy, water, transportation, and other physical-consequence sectors. NIST explicitly noted that critical infrastructure will increasingly rely on AI across information technology, operational technology, and industrial control systems, and that this convergence requires specialized risk management approaches beyond those the baseline AI RMF provides.

The development of a critical infrastructure-specific AI RMF profile signals that federal risk management thinking has begun to catch up with the concentration reality. The profile's Community of Interest invitation, which NIST issued alongside the concept note, sought participation from across the entire critical infrastructure ecosystem—all sectors, organizational roles, and supply chain partners. The supply chain emphasis is notable: it reflects an awareness that the risks to critical infrastructure AI deployments originate not only in the organizations deploying AI but in the vendor supply chains upon which those deployments depend.

Congressional Activity and the Critical Infrastructure Designation Debate

The question of whether AI data centers should formally qualify as critical infrastructure under U.S. law reached congressional hearings in May 2026 [8]. The House Homeland Security Subcommittee on Cybersecurity and Infrastructure Protection received testimony on the issue, with Representative Andy Ogles observing that the existing federal framework provides no clear, unified approach to data center security and does not designate which federal agency bears primary responsibility for their protection. The

Data Infrastructure Risk Reduction Act (H.R. 8711) was introduced to address this gap [23], requiring the identification of data centers that should be treated as critical infrastructure and mandating a congressional strategy for defending them against external breaches.

The formal critical infrastructure designation debate matters for the dual-role problem because designation would trigger sector-specific regulatory obligations, information-sharing requirements, and federal coordination mechanisms that currently do not apply to AI vendors. Regulated critical infrastructure sectors in the United States are required to maintain sector-specific security plans, participate in federal information sharing programs, and meet resiliency standards defined through sector coordinating councils. Applying analogous requirements to AI infrastructure vendors would create accountability mechanisms for vendor security practices that current software and cloud service provider regulations do not provide.

Government AI Security Oversight

The Center for AI Safety and Innovation announced evaluation partnerships with Google DeepMind, Microsoft, and Elon Musk's xAI in May 2026, building on earlier agreements with OpenAI and Anthropic, to conduct pre-deployment security evaluations of frontier AI capabilities [18]. The Trump administration simultaneously took a more interventionist approach to AI security oversight than its predecessor, testing major AI providers through government-affiliated evaluation bodies and taking a posture in which national security considerations shaped which AI vendors would receive access to classified government networks. The Department of Defense announced agreements with eight technology companies for classified AI deployments in early May 2026, with Anthropic explicitly excluded over disagreements about safety guardrails—an episode that illustrated how government reliance on specific AI vendors for national security functions amplifies the consequences of vendor relationships and creates incentive structures that may not align with systemic risk management [19].

Recommendations for Enterprise Risk Management

Establishing AI Vendor Concentration Limits

Enterprises should establish explicit concentration thresholds for AI vendor dependency, analogous to the third-party concentration limits already common in financial services risk management. A practical approach involves mapping all AI-dependent workflows across three categories: business-critical workflows where AI model failure would halt operations; security-dependent workflows where AI model integrity is required for threat detection and response; and governance workflows where AI outputs inform compliance or regulatory reporting. For each category, an enterprise should assess the degree to which a single vendor compromise

could simultaneously affect multiple workflows, and set ceiling thresholds—such as no single vendor supplying more than forty percent of security-dependent AI capabilities—that trigger diversification requirements when exceeded.

This concentration mapping exercise will typically reveal that the informal default of routing all AI workloads through a single hyperscaler's bundled stack has produced concentration levels that no risk officer would have formally approved. The vertical integration strategies of major AI vendors have made concentration the path of least resistance; managing it requires deliberate policy intervention rather than organic vendor management.

Separating Security Tooling from Business AI Vendors

Wherever operationally feasible, enterprises should resist acquiring AI-powered security tooling from the same vendors supplying their primary business AI capabilities. This separation principle is designed to address the dual-role correlation risk: when a single vendor supplies both the business AI platform and the security detection system, the detection system's integrity is necessarily dependent on the vendor's security posture. The separation is not a perfect mitigation—the supply chain interdependencies discussed earlier mean that competing vendors may share infrastructure or library dependencies—but it reduces the probability that a single vendor compromise simultaneously degrades both business capabilities and the ability to detect that compromise.

Where full vendor separation is impractical, enterprises should implement compensating controls: maintaining at least one security detection capability that operates independently of AI model inference (rule-based detection, signature matching, or behavior analytics that does not depend on a large language model for classification), and establishing out-of-band incident notification mechanisms that do not route through AI-mediated communication channels during incident response.

Requiring Transparency and Incident Notification

Enterprises with significant AI vendor dependencies should negotiate contractual requirements for timely incident disclosure, supply chain compromise notification, and model behavioral change notification as conditions of vendor agreements. The current state of AI vendor contracts often contains broad indemnification language protecting the vendor from liability for model behavioral changes, with no affirmative disclosure obligation when the vendor identifies security incidents affecting models in production use. This disclosure asymmetry is not acceptable for security-critical applications.

Contract terms should specifically address: notification timelines for security incidents affecting models in production (a 72-hour notification window is consistent with GDPR breach notification standards and represents a reasonable baseline); disclosure of known vulnerabilities in third-party dependencies that are part of the model serving stack; and advance notification of model updates that may alter behavioral

characteristics relevant to the enterprise's use case. Enterprises should also seek audit rights—even if exercised only periodically through third-party assessors—that provide some visibility into vendor security practices at the infrastructure layer.

Preparing for Vendor Isolation Scenarios

CISA's CI Fortify initiative established the operational principle that critical infrastructure operators should develop and test plans for continuing operations when third-party dependencies, including AI vendors, are unavailable [22]. Enterprises across all sectors should apply this principle to their AI-dependent workflows. The planning exercise involves identifying the minimum viable operational state achievable without any AI vendor connectivity, documenting the manual or rule-based alternatives for each AI-dependent process, and testing those alternatives periodically so that they remain operationally viable rather than theoretical fallbacks.

The isolation planning exercise is valuable not only for resilience purposes but as a risk visibility mechanism. Organizations that conduct the exercise typically discover that they have far fewer fallback capabilities than they assumed, that several critical workflows have no viable non-AI alternative, and that the operational cost of AI vendor isolation is much higher than the vendor dependency maps suggested. This discovery is itself important risk management information that should inform concentration limit decisions and vendor contract negotiations.

Participating in Sector-Wide Information Sharing

The defense against systemic AI concentration risk cannot be achieved by individual enterprises acting in isolation; it requires sector-level coordination on incident intelligence, supply chain vulnerability information, and model behavioral anomaly data. Enterprises should participate actively in sector-specific information sharing and analysis centers (ISACs) and in cross-sector initiatives such as the Coalition for Secure AI, which is developing shared standards and best practices for AI security across more than 45 member organizations [10]. When an AI vendor incident occurs, early information sharing within a sector allows organizations using the same vendor to detect whether they are affected, implement compensating controls, and coordinate on disclosure advocacy to the vendor—leveraging the collective bargaining power of the enterprise customer base in ways that individual organizations cannot.

Conclusions

The AI developer ecosystem has, in a remarkably short period, acquired the structural characteristics that have historically defined critical infrastructure: concentrated ownership, broad societal dependency, limited short-term substitutability, and cascading failure potential when a component fails. The additional dimension that distinguishes AI infrastructure from previous critical infrastructure sectors is the dual-role problem: the vendors who hold the most valuable targets for sophisticated adversaries are simultaneously the primary providers of the defensive tools enterprises use against those adversaries.

This dual role is not the result of malicious intent or inadequate planning. It reflects the natural market dynamics of a technology sector where the same capabilities that make AI models useful for business applications also make them useful for security applications, and where the capital intensity of frontier model development concentrates both categories of capability with the same small number of providers. But structural neutrality on the part of the vendors does not reduce the risk to enterprises and governments that depend on them. The risk is systemic and structural, and it will not be addressed by better vendor management practices alone.

The policy and regulatory response is beginning to catch up with this reality. The NIST AI RMF Critical Infrastructure Profile, the congressional hearings on data center designation, the OSTP memorandum on adversarial distillation, and CISA's CI Fortify initiative all reflect a growing recognition that AI vendor concentration requires governance mechanisms analogous to those applied to utilities, financial market infrastructure, and telecommunications. The pace of this regulatory evolution will need to accelerate. Adversaries—including nation-state actors with documented operational programs targeting U.S. AI infrastructure—are not waiting for the policy framework to mature.

Enterprises should not wait either. The concentration mapping, vendor separation, contract transparency, isolation planning, and information-sharing practices described in this paper are actionable now, within existing legal and regulatory frameworks, and represent the minimum viable risk management response to a concentration exposure that has already materialized into documented incidents. Organizations that delay because the regulatory mandate has not yet arrived are making a gamble against a threat landscape that has already demonstrated the willingness and capability to exploit the defender-target paradox at scale.

CSA Resource Alignment

The risks documented in this paper intersect directly with several active CSA frameworks and initiatives, which provide implementation guidance for the recommendations offered here.

The **AI Controls Matrix (AICM)** provides a comprehensive control framework for AI security governance that maps across multiple organizational roles: AI customers, application providers, cloud service providers, model providers, and orchestrated service providers. The concentration risk management recommendations in this paper—particularly contract transparency requirements, supply chain visibility obligations, and incident notification standards—align with the AICM's model provider and orchestrated service provider control domains. Enterprises seeking to operationalize concentration limits should use the AICM as the control baseline for assessing vendor compliance. The AICM is a superset of the Cloud Controls Matrix (CCM), and organizations already using CCM for cloud vendor assessment can extend their existing assessment processes to cover AI-specific control domains through AICM adoption.

The **MAESTRO framework** (Agentic AI Threat Modeling) directly addresses the attack surface inheritance risks described in this paper. MAESTRO's seven-layer threat model covers physical compute, data operations, AI models, agent frameworks, deployment, security monitoring, and governance—each of which represents a potential site of dual-role compromise in the vendor concentration scenarios analyzed here. The behavioral misalignment and supply chain injection attack vectors described in this paper correspond specifically to MAESTRO's AI model layer and agent framework layer threat categories. Enterprises implementing MAESTRO threat modeling for their AI agent deployments will find that the vendor concentration issue surfaces naturally as a systemic risk factor in the governance layer.

The **CSA AI Model Risk Management Framework** [21] provides methodological guidance on Model Cards, Risk Cards, and Scenario Planning that supports the isolation planning and vendor behavioral monitoring recommendations in this paper. The framework's Scenario Planning pillar—which addresses stress testing AI systems against adversarial and failure scenarios—provides a structured methodology for the AI vendor isolation exercises recommended above. Risk Cards, which document AI-specific risk assessments at the model level, provide a mechanism for capturing dual-role concentration risk as a model-level concern that persists through vendor contract cycles.

The **CSA AI Organizational Responsibilities** publication series [20]—covering governance and risk management, core security responsibilities, and AI tools and applications—addresses the organizational accountability structures needed to manage the concentration risks described in this paper. Specifically, the governance guidance addresses how enterprises should allocate responsibility for AI vendor risk between security, procurement, legal, and AI operations teams. The current common practice of treating AI vendor relationships primarily as procurement decisions, with security assessment as a secondary review step, is inadequate for managing the dual-role concentration risk identified here. CSA's guidance supports the elevation of AI vendor security assessment to a primary governance concern.

The **STAR for AI** registry program enables enterprises to assess and disclose their AI vendor relationships and security posture through a standardized self-assessment framework. As AI vendor concentration risk becomes a standard component of enterprise AI governance, STAR for AI assessments will provide a mechanism for enterprises to communicate their concentration exposure to customers, regulators, and auditors in a standardized format.

References

[1] Synergy Research Group. "Cloud Market Share Trends – Big Three Together Hold 63% While Oracle and the Neoclouds Inch Higher." Synergy Research Group, Q1 2026.

[2] Constellation Research. "Google Cloud, AWS, Microsoft Azure: The AI Vertical Integration Race." Constellation Research, 2026.

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

[4] Nextgov/FCW. "White House Accuses China of 'Deliberate, Industrial-Scale Campaigns' to Steal US AI Models." Nextgov, April 2026.

[5] Security Boulevard. "What the Latest OpenAI Security Breach Reveals About the State of AI Protection." Security Boulevard, December 2025.

[6] Outlook Business. "OpenAI and Anthropic's Data Supplier Was Hacked—Here's What We Know." Outlook Business, 2025.

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

[8] ClearanceJobs. "Congress Questions Whether AI Data Centers Should Be Declared Critical Infrastructure." ClearanceJobs, May 6, 2026.

[9] Yahoo Finance. "Hyperscalers Hit $700 Billion in 2026 AI Spending Plans." Yahoo Finance, 2026.

[10] Coalition for Secure AI. "Addressing What's Next in Securing Enterprise AI." CoSAI, 2026.

[11] Hashrateindex. "Inside the Custom AI Chip Race: Google, AWS, Microsoft, Meta, OpenAI." Hashrate Index, 2026.

[12] IDC. "Agentic AI Governance: When AI Becomes Critical Infrastructure." IDC, 2026.

[13] CISA. "Principles for the Secure Integration of Artificial Intelligence in Operational Technology." CISA, December 2025.

[14] DevOps.com. "OpenAI's Daybreak Challenges Anthropic in AI Cybersecurity Race." DevOps.com, 2026.

[15] SecureWorld. "The Dual-Front War: Navigating AI as Both Engine and Target." SecureWorld, 2026.

[16] Dark Reading. "Microsoft & Anthropic MCP Servers at Risk of RCE, Cloud Takeovers." Dark Reading, 2025.

[17] NIST. "Concept Note: AI RMF Profile on Trustworthy AI in Critical Infrastructure." NIST, April 7, 2026.

[18] CNBC. "Trump Admin Moves Further Into AI Oversight, Will Test Google, Microsoft and xAI Models." CNBC, May 5, 2026.

[19] Center for American Progress. "The Department of Defense's Conflict With Anthropic and Deal With OpenAI Are a Call for Congress To Act." CAP, 2026.

[20] CSA. "AI Organizational Responsibilities: Core Security Responsibilities." Cloud Security Alliance, 2024.

[21] CSA. "AI Model Risk Management Framework." Cloud Security Alliance, 2024.

[22] CISA. "CISA Unveils New Initiative to Fortify America's Critical Infrastructure." CISA, May 5, 2026.

[23] U.S. Congress. "Data Infrastructure Risk Reduction Act, H.R. 8711." 119th Congress, 2026.

[24] Built In. "How Circular Financing Is Fueling the AI Boom." Built In, 2025.

[25] TechRepublic. "NVIDIA Blackwell GPUs Sold Out: Demand Surges, What's Next?." TechRepublic, 2025.