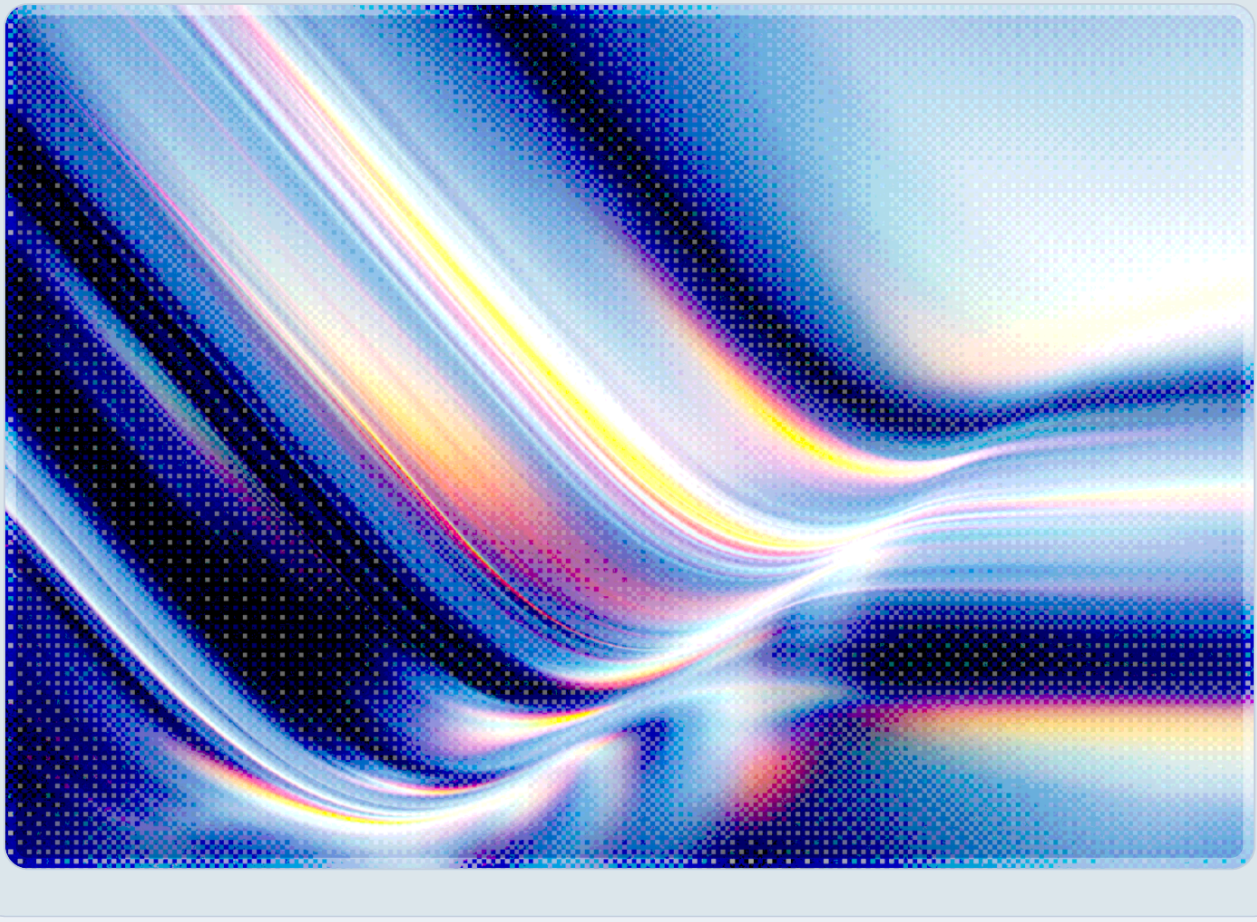
CSAI Foundation | Cloud Security Alliance

AI Compute Concentration and Systemic Risk

Enterprise Vulnerabilities in the Hyperscaler-AI Provider Oligopoly

2026-05-09

AI-assisted Rapid Research



CSAI

cloud security alliance®

Table of Contents

Executive Summary

The rapid commercialization of artificial intelligence has produced a market structure that bears an uncomfortable resemblance to the conditions regulators spent decades trying to prevent in telecommunications and financial services: a small number of vertically integrated providers controlling the full stack from raw compute through model inference and application delivery. Today, three hyperscalers — Amazon Web Services, Microsoft Azure, and Google Cloud — account for the overwhelming majority of AI-grade infrastructure capacity available to enterprises, while a single chip designer, NVIDIA, controls approximately 80 to 90 percent of the AI accelerator market by revenue [1][2]. The AI application layer is similarly concentrated, with a handful of model providers, most of them dependent on the same hyperscalers for their own compute, serving the bulk of enterprise inference traffic.

This structure creates systemic risks that extend well beyond traditional vendor risk. When enterprises embed AI deeply into revenue-generating processes, and when those AI services depend on a chain of concentrated providers, the failure, disruption, or adversarial compromise of any link in that chain propagates broadly. The documented outage patterns of 2024 through early 2026 — during which both OpenAI and Anthropic struggled to maintain 99 percent availability, resulting in effectively more than three and a half days of annual downtime — illustrate the practical consequences of this fragility [3]. At the infrastructure layer, hyperscaler outages have taken down entire categories of AI services simultaneously, including scenarios where a Cloudflare disruption in November 2025 impacted not only AI APIs but downstream services relying on them [4].

The regulatory community has recognized these dynamics. The European Union's AI Act establishes a systemic risk classification for general-purpose AI models based on training compute thresholds, with corresponding obligations and fines of up to three percent of worldwide annual turnover [5]. The U.S. Federal Trade Commission has examined partnerships between hyperscalers and AI model developers, identifying concerns about access to compute, switching costs, and competitive entrenchment [6]. The UK's Competition and Markets Authority identified six firms — Google, Amazon, Microsoft, Meta, Apple, and NVIDIA — as controlling the foundation model value chain in ways that could distort competition [7].

This whitepaper analyzes the anatomy of AI compute concentration, its enterprise security implications, the evolving regulatory response, and concrete steps organizations can take to reduce systemic exposure. The goal is not to argue against AI adoption, but to ensure that the security and risk communities understand the structural vulnerabilities embedded in the current market and act accordingly before a major cascading incident makes the stakes undeniable.

1. Introduction and Background

The story of enterprise technology has always been a story of concentration followed by disruption followed by re-concentration. Mainframe dominance gave way to distributed computing, which gave way to the personal computer era, which gave way to cloud consolidation. Each transition promised democratization and delivered a new oligopoly. The AI era is following the same arc, but compressing the timeline in ways that outpace institutional risk management.

Enterprises are integrating AI capabilities into customer-facing products, internal operations, and critical decision-making processes at a pace that would have seemed implausible five years ago. A 2025 Gartner analysis projected worldwide AI spending would total nearly $1.5 trillion in 2025, with IT services companies and hyperscalers accounting for over 70 percent of that spending [8]. This integration is not superficial. Organizations are embedding AI into revenue operations, supply chain logistics, compliance processes, fraud detection, and clinical decision support — functions where degraded or unavailable AI services translate directly into operational and financial harm.

The speed of this integration has outrun the development of risk management frameworks adequate to the concentrated market structure underlying it. Traditional vendor risk management practices, calibrated for software vendors or data processors, are poorly suited to a supply chain in which a single provider controls raw compute, provides the model, hosts the API, and also operates the enterprise's underlying cloud infrastructure. The simultaneous failure modes this creates are not adequately captured by standard third-party risk assessments.

Understanding these risks begins with understanding the market structure that created them. The concentration operating today across the AI stack — from semiconductor fabrication through model inference — is not incidental. It is the product of capital dynamics, regulatory lag, and technical barriers to entry that mutually reinforce one another. Addressing the risk requires understanding each layer.

2. The Anatomy of AI Compute Concentration

2.1 Semiconductor Fabrication: The Deepest Chokepoint

Before a single AI model can be trained or a single inference served, the underlying compute must exist as physical silicon. The production of the most advanced AI accelerators is concentrated in a single manufacturing facility — Taiwan Semiconductor Manufacturing Company — which produces approximately 90 percent of the world's most advanced process-node chips [9]. This is not merely a market share figure; it represents a single geographic concentration of irreplaceable manufacturing capability that underpins the entire global AI buildout. A disruption to TSMC, whether from geopolitical conflict, natural disaster, or operational failure, would constrain AI compute supply globally in ways that no hyperscaler investment could rapidly remedy.

At the design layer, NVIDIA controls between 80 and 90 percent of the AI accelerator market by revenue, generating over $100 billion annually as of 2025 [2]. The company's dominance extends beyond its hardware to the CUDA software ecosystem, which represents a decade of developer investment and creates substantial switching costs even as competing hardware becomes technically capable. AMD holds approximately eight percent of the discrete GPU market, and Intel's AI accelerator presence remains below one percent [1]. Custom silicon from hyperscalers — Google's Tensor Processing Units, Amazon's Trainium, and comparable investments from Meta and Microsoft — is growing at an estimated 44.6 percent compound annual growth rate and offers 40 to 65 percent total cost of ownership advantages over merchant silicon for internal workloads [10]. However, these advantages accrue primarily to the hyperscalers themselves; enterprise customers accessing AI through cloud APIs obtain no direct exposure to custom silicon economics.

2.2 Cloud Infrastructure: Three-Player Market for AI Compute

Above the silicon layer, the cloud infrastructure market for AI workloads is effectively a three-player market. Amazon Web Services, Microsoft Azure, and Google Cloud collectively control the overwhelming majority of AI-grade compute available to enterprises. North America controlled 44.7 percent of global AI-ready hyperscale data center infrastructure capacity in 2025, with those three providers dominating that share [11]. Their capital deployment reflects the stakes: AWS has committed to approximately $200 billion in infrastructure investment focused on AI capacity expansion, while Meta has committed $125 to $145 billion annually for data centers, networking, and GPU clusters [12].

The revenue growth figures underscore the demand dynamics. Google Cloud reported 63 percent year-over-year growth in Q1 2026 with a $460 billion order backlog; Azure reported 40 percent growth; AWS reported 28 percent growth [11]. These are not the growth curves of a competitive market reaching equilibrium — they are the curves of a market where demand consistently outstrips available supply, and where entry barriers prevent the supply response that would normally moderate both prices and concentration.

2.3 The AI Model Layer: Concentration by Dependency

The foundation model layer presents a different but equally concerning form of concentration. A small number of model providers — OpenAI, Anthropic, Google DeepMind, and Meta — account for the majority of commercially deployed general-purpose AI capabilities. What makes this layer structurally significant, beyond simple market share, is the dependency relationship: model providers depend on hyperscalers for their own training and inference compute, and hyperscalers have responded by making strategic investments in model providers, creating partnership structures the FTC has examined for competitive effects [6].

The result is a set of interlocking dependencies that enterprise buyers navigate without full visibility. An enterprise using a model provider's API may not realize that the model provider's infrastructure is hosted entirely on a single hyperscaler, or that the hyperscaler has contractual rights to model architectures and training data through its investment structure. These are not merely business relationships — they are structural dependencies that shape the risk profile of the enterprise's AI supply chain whether or not the enterprise knows they exist.

Layer	Dominant Players	Concentration Indicator
Chip fabrication	TSMC	~90% of advanced-node chips
AI accelerator design	NVIDIA	80-90% market share by revenue
Cloud infrastructure (AI)	AWS, Azure, Google Cloud	~3-player market for enterprise AI
Foundation models	OpenAI, Anthropic, Google, Meta	Few providers; hyperscaler-dependent
AI application APIs	Same as above, plus embedded services	Oligopoly with switching costs

3. Enterprise Vulnerabilities in a Concentrated Market

3.1 Reliability Risk and Documented Outage Patterns

The most immediate and measurable vulnerability is availability. When the same providers serve a large portion of enterprise AI workloads, provider-level outages become sector-level events. The record of the past two years illustrates this clearly. OpenAI experienced a global ChatGPT outage in May 2024, followed by further significant disruptions in June 2024, a 15-hour outage in June 2025 affecting ChatGPT, Sora, and API services, and a major December 2025 incident [13]. Anthropic's Claude service experienced elevated error rates beginning in March 2026, with a second outage occurring less than 24 hours after recovery — part of a pattern of repeated service disruptions documented across the same period [14]. The aggregate implication is that neither leading model provider maintained 99 percent annual availability — translating to more than three and a half days of downtime per year in expectation [3].

At the infrastructure layer, concentration compounds these risks. A November 2025 Cloudflare outage cascaded through AI services including OpenAI and others, affecting an estimated billions of global users [4]. An October 2025 AWS outage took down AI applications hosted on its infrastructure simultaneously, demonstrating that hyperscaler failures affect not individual enterprises but entire populations of dependent services at once [15]. The financial exposure from these events is substantial: industry research finds that over 90 percent of mid-size and large enterprises report that a single hour of downtime costs more than $300,000, with nearly half estimating losses in the range of $1 million to $5 million per hour [27].

3.2 Vendor Lock-in and Operational Dependency

The security implications of vendor lock-in in AI extend considerably beyond the familiar concerns of software portability. When an enterprise embeds a specific model provider's API into automated customer support, revenue operations, or supply chain decision systems, it creates a dependency that is not easily reversed. Model providers can change terms of service, deprecate API versions, alter model behavior through silent retraining, or modify guardrails in ways that break enterprise workflows without notice. These are not hypothetical concerns — they represent patterns that have occurred across major AI providers in the 2024-2026 period.

The lock-in operates at multiple levels simultaneously. At the compute level, enterprises often build on cloud-native AI services that are coupled to specific hyperscaler infrastructure, creating dependencies on proprietary networking, storage, and identity systems that resist portability. At the model level, prompt engineering, fine-tuning investments, and integration architectures developed for one model's capabilities and behavior often do not transfer cleanly to alternative models, even those marketed as similar. At the data

level, training data uploaded to hosted fine-tuning services may become subject to the provider's data retention and usage policies in ways that complicate migration. The accumulation of these dependencies across organizational AI deployments creates a risk exposure that is rarely inventoried or assessed holistically.

3.3 AI Supply Chain Opacity

Traditional third-party risk management frameworks evaluate vendors based on their data processing practices, security controls, and contractual obligations. These frameworks are inadequate for AI supply chains, which introduce categories of risk that do not fit the standard assessment template. AI models learn from data, evolve through retraining cycles, and can drift into behaviors that are materially different from their evaluated state — all without triggering the change management notifications that would prompt a third-party risk review under conventional frameworks [17].

Organizations frequently lack visibility into the full depth of their AI supply chains. An enterprise may assess its primary AI vendor but remain unaware that the vendor's model was trained on data from a third-party scraping operation, that inference is served through a subprocessor with different data handling practices, or that the underlying compute is hosted in a jurisdiction with regulatory implications for the enterprise's own compliance posture. Research on AI supply chain security identifies this blindness as a critical gap: organizations neglect the third-party AI vendors integral to their ecosystem because their risk management tools were designed for software vendors, not for systems that learn, adapt, and interact with enterprise data dynamically [17][18].

3.4 Algorithmic Herding and Correlated Failure

Beyond the operational risks, AI market concentration creates a subtler but potentially more consequential category of systemic risk: the tendency toward correlated behavior when large populations of enterprises rely on the same models making decisions about similar data in similar contexts. This risk has received the most formal analysis in financial services, where regulators have raised concerns about AI "monoculture." Former SEC Chair Gary Gensler warned explicitly that deep learning's characteristics could drive convergence on a small number of dominant data and model providers, resulting in synchronized investment strategies and the amplification of market volatility [19]. Academic modeling of this effect suggests tail-loss amplification of 18 to 54 percent in financial markets under AI monoculture conditions — a magnitude that is economically significant relative to Basel III countercyclical capital buffers [20][28].

The dynamics are not unique to financial services. Enterprises in any sector that rely on the same AI models to classify risk, prioritize resources, or make allocation decisions may find themselves responding to market conditions or operational signals in correlated ways that they individually would not choose and collectively

cannot easily avoid. When AI market structure determines model access, the market structure itself becomes a systemic risk factor.

Vulnerability Category	Mechanism	Enterprise Impact
Availability concentration	Provider outage cascades across dependent services	Revenue loss; operational disruption
Vendor lock-in	Migration barriers prevent response to provider failure or terms change	Strategic inflexibility; operational fragility
Supply chain opacity	Multi-tier dependencies not visible to enterprise	Unknown compliance, data, and security exposure
Algorithmic herding	Correlated model outputs across enterprises using same providers	Amplified sector-level volatility and correlated failures
Geographic concentration	TSMC manufactures ~90% of advanced chips in a single region	Long-duration compute scarcity risk from geopolitical or physical events

4. The Regulatory Response

4.1 European Union: Systemic Risk Classification

The EU AI Act, which entered into force in August 2024, introduces a regulatory category specifically designed to address the risk profile of the most powerful general-purpose AI models. Under Articles 51 through 55 of the Act, a general-purpose AI model is presumed to present systemic risk when the cumulative computing power used for its training exceeds 10^25 floating-point operations [5]. This threshold is designed to capture models whose scale gives them the potential to affect broad populations and sectors simultaneously — precisely the dynamic that compute concentration enables.

Providers of models classified as presenting systemic risk are subject to a distinct set of obligations: conducting model evaluations and adversarial testing before and after deployment, tracking and reporting serious incidents to the European AI Office and national authorities, maintaining cybersecurity protections commensurate with their risk profile, and publishing technical documentation and training content summaries [21][22]. Non-compliance is subject to fines of up to three percent of annual worldwide turnover or 15 million euros, whichever is higher [21]. The Act's systemic risk provisions represent the most direct regulatory acknowledgment to date that AI market concentration creates risks requiring intervention beyond standard product liability or data protection frameworks.

4.2 United States: Antitrust and Competition Oversight

The U.S. regulatory response has unfolded through existing competition law enforcement rather than AI-specific legislation. The FTC issued a staff report examining partnerships between major cloud providers — Alphabet, Amazon, and Microsoft — and AI developers, including Anthropic and OpenAI. The report identified specific competitive concerns: that these partnerships may shape access to computing resources and engineering talent in ways that disadvantage competitors, that they increase switching costs for AI developer partners, and that they provide cloud service providers access to sensitive technical and business information unavailable to competitors [6]. The FTC also finalized new Hart-Scott-Rodino rules in 2025 that close loopholes allowing talent-based acquisitions and licensing deals to transfer effective control of company assets without triggering merger review — a reform specifically relevant to the pattern of hyperscaler investment in AI model developers [23].

The GAO, in a May 2025 report (GAO-25-107197), warned that financial instability may arise from reliance on concentrated third-party AI service providers, noting that a failure at one provider could affect an outsized number of financial companies and thereby increase systemic risk [26]. The Commodity Futures Trading Commission echoed these concerns in June 2025, identifying growing cybersecurity risks amplified

by AI provider concentration [24]. Congressional attention has followed, with hearings in the House Financial Services Committee and House Judiciary Subcommittee examining AI competition and market stability in the context of financial services [25].

4.3 United Kingdom: Foundation Model Review

The UK Competition and Markets Authority conducted a foundational review of the AI foundation model market that produced the clearest structural analysis of concentration risks from any regulatory body. The CMA identified three distinct competition risk mechanisms arising from the dominance of six firms — Google, Amazon, Microsoft, Meta, Apple, and NVIDIA — in the AI value chain [7]. First, input control risk: these firms' control over compute resources, proprietary training data, and specialized engineering talent may prevent competitors from accessing these inputs at the terms required to develop competitive foundation models. Second, market position exploitation: existing dominance in mobile platforms, search, and productivity software gives incumbent firms structural advantages in distributing AI services that competitors cannot replicate. Third, entrenchment through partnerships: strategic investments by incumbent firms in AI developers may reduce the developers' incentive and ability to serve as competitive alternatives, converting potential competitors into supply chain dependents [7].

The CMA's analysis is significant for enterprise security practitioners because it describes not just competitive dynamics but the structural conditions that create systemic risk. A market in which a small number of firms control all viable supply chain inputs, all major distribution channels, and the primary model developers simultaneously is a market that is structurally fragile regardless of each individual firm's operational quality.

4.4 Regulatory Convergence and Remaining Gaps

Despite the breadth of regulatory attention, significant gaps remain. No major jurisdiction has yet established mandatory concentration limits for AI compute or model provision, or required enterprises to demonstrate multi-provider resilience as a condition of deploying AI in critical operations. The EU AI Act's systemic risk provisions address model providers but do not directly regulate the infrastructure concentration that enables them. U.S. competition enforcement operates reactively through merger review and investigation rather than through prospective structural requirements. The UK CMA's review has identified risks but has not yet produced binding remedies.

The net effect is that enterprises operating in 2026 cannot rely on regulatory frameworks to manage the risks of AI compute concentration on their behalf. The responsibility for identifying, measuring, and mitigating these risks falls to enterprise security, risk, and technology leadership — operating in a regulatory environment that is still developing the tools to address the problem systematically.

5. Recommendations for Enterprise Security and Resilience

5.1 Build and Maintain an AI Supply Chain Inventory

The precondition for managing AI concentration risk is visibility. Enterprises should maintain a current inventory of every AI service embedded in their operations, mapping each service to its model provider, the compute infrastructure the provider uses, and the geographic and contractual dependencies that underlie that infrastructure. This is not a one-time exercise: AI supply chains change as providers retrain models, renegotiate hyperscaler relationships, or change API structures, and the inventory must be actively maintained to remain useful.

The inventory should extend beyond direct AI API consumption to include AI capabilities embedded in purchased software products, platforms, and managed services. Many enterprise AI dependencies are indirect — surfacing through a vendor's product rather than through a directly contracted AI service — and these are precisely the dependencies most likely to be missed in a conventional third-party risk review.

5.2 Establish Multi-Provider Resilience Architecture

Enterprises that have embedded AI into critical operations should develop and test contingency architectures that allow continued operation when a primary AI provider is unavailable. This requires more than identifying an alternative provider in advance; it requires confirming that the alternative provider's model behavior is sufficiently similar to maintain acceptable output quality for the relevant use cases, that integration architecture supports provider switching without extensive re-engineering, and that the organization has practiced the switch under realistic conditions.

For applications where AI unavailability would halt critical operations, organizations should consider maintaining production-grade integrations with at least two providers simultaneously, balancing traffic across both during normal operations to maintain operational familiarity and avoid the cold-start problems that plague seldom-exercised fallback systems. The cost of this redundancy should be evaluated against the documented cost of AI-related downtime: industry research finds that over 90 percent of mid-size and large enterprises report that a single hour of downtime costs more than $300,000, with nearly half estimating losses in the range of $1 million to $5 million per hour [27].

5.3 Apply Concentration Limits in Procurement

Organizations should establish explicit policies limiting the share of critical AI workloads that can be served by any single provider or infrastructure chain. These policies should account for the depth of concentration, not just the surface provider: an enterprise that uses multiple model providers but whose primary and backup model providers both run exclusively on the same hyperscaler has achieved less diversification than its vendor roster suggests.

Procurement requirements should include contractual provisions addressing provider obligations around advance notice for API deprecations and model behavioral changes, data portability and export rights, incident notification timelines, and transparency about subprocessors and infrastructure dependencies. Where providers decline to make these commitments, this should be treated as a risk factor in supplier selection.

5.4 Treat AI Infrastructure as Critical Infrastructure

The operational integration of AI into revenue-generating and compliance-critical processes makes AI infrastructure functionally equivalent to other categories of critical infrastructure in many organizations, even where it is not legally recognized as such. Enterprise risk frameworks should reflect this by applying equivalent resilience requirements to AI services as they would to core ERP, financial systems, or operational technology. This includes defining recovery time objectives and recovery point objectives for AI services, testing recovery procedures regularly, and including AI service unavailability scenarios in business continuity planning.

Organizations in regulated industries — particularly financial services, healthcare, and critical infrastructure sectors — should proactively engage with regulators on how AI concentration risk fits within existing operational resilience requirements, anticipating that regulatory guidance in this area will become more prescriptive as AI integration deepens and outage incidents accumulate.

5.5 Monitor the Evolving Competitive Landscape

The semiconductor and AI infrastructure landscape is evolving rapidly. Hyperscaler custom silicon is growing at 44.6 percent CAGR and offers 40 to 65 percent TCO advantages for internal workloads [10]. NVIDIA's inference market share is projected to decline from its current 80 to 90 percent range as alternative architectures mature [2]. Open-weight model alternatives from Meta's Llama series, Mistral, and others provide options for enterprises willing to operate their own model infrastructure, reducing dependency on hosted model providers. Geopolitically motivated diversification of advanced semiconductor manufacturing — including TSMC expansion in Arizona and semiconductor investment programs in Japan and Europe — may gradually reduce the geographic concentration risk over multi-year horizons [9].

Enterprises should track these developments as part of their AI risk management discipline, reassessing concentration exposure annually and adjusting architecture and procurement practices as viable alternatives become available at the scale and reliability their operations require.

6. CSA Resource Alignment

The risks described in this whitepaper sit at the intersection of several existing CSA frameworks and research streams, providing enterprise practitioners with a structured foundation for implementing the recommendations above.

6.1 AI Controls Matrix

CSA's AI Controls Matrix (AICM v1.0) provides the most directly applicable framework for managing AI supply chain and provider concentration risks. The AICM's 18 control domains explicitly address AI supply chain security, shared security responsibility modeling for multi-tier AI supply chains, and governance requirements for third-party AI providers. Enterprises implementing the AICM's auditing guidelines for AI customers and orchestrated service providers will find specific control structures relevant to the inventory, procurement, and resilience requirements described in Section 5 of this whitepaper. The AICM's Shared Security Responsibility Model is particularly useful for mapping accountability across the layered dependencies — from hyperscaler infrastructure through model provider through enterprise application — that define AI supply chain risk in a concentrated market.

6.2 MAESTRO Framework

CSA's MAESTRO framework for agentic AI threat modeling provides a structured methodology for identifying how AI concentration risks surface in agentic architectures specifically. As enterprises deploy AI agents with autonomy over consequential actions — executing transactions, modifying configurations, communicating with customers — the resilience and behavioral consistency of the underlying model provider becomes a safety concern, not merely an operational one. MAESTRO's threat modeling approach helps organizations identify where agentic AI dependencies on specific providers create unacceptable risk concentrations, and supports the design of human oversight mechanisms that can detect and interrupt problematic behavior when model providers silently alter model outputs through retraining.

6.3 Cloud Controls Matrix and STAR

CSA's Cloud Controls Matrix (CCM) provides controls relevant to the cloud infrastructure concentration dimension of AI provider risk, particularly in the supply chain management, business continuity, and third-party management domains. Organizations using the STAR (Security Trust Assurance and Risk) program to evaluate cloud service providers should extend their STAR-based assessments to include the AI-specific

supply chain questions that the concentrated AI market raises — including questions about providers' own hyperscaler dependencies and about the depth of integration between model providers and the cloud infrastructure providers that invest in them.

For enterprises pursuing STAR Level 2 assessments of AI providers, the concentration analysis framework described in Section 2 of this whitepaper provides a useful complement to standard STAR audit procedures, helping auditors identify structural risks that STAR's control-by-control review might not surface on its own.

6.4 Zero Trust Guidance

CSA's Zero Trust architecture guidance applies to AI provider relationships through its core principle that no provider should be implicitly trusted simply because it has been previously onboarded or certified. The documentation of silent model behavioral changes, API deprecations without adequate notice, and terms-of-service modifications that retroactively affect data handling all point to the need for continuous verification of AI provider behavior rather than reliance on point-in-time assessments. Zero Trust principles applied to AI supply chains suggest treating each model inference as a transaction that should be validated against expected behavioral parameters — a practice that becomes more operationally significant as AI outputs drive consequential decisions.

7. Conclusions

The concentration of AI compute, infrastructure, and model provision in a small number of vertically integrated providers is not a temporary feature of an immature market that competition will naturally resolve. The capital requirements for competitive entry, the CUDA ecosystem's switching costs, the geographic concentration of advanced semiconductor manufacturing, and the structural advantages incumbent hyperscalers enjoy in bundling AI services with existing enterprise relationships all suggest that concentration will persist and may deepen in the near term, even as specific players within the oligopoly compete vigorously against one another.

For enterprise security practitioners, the implication is that AI concentration risk is a durable feature of the landscape requiring systematic management rather than a temporary condition to be waited out. The regulatory frameworks being developed across the EU, UK, and U.S. will eventually establish minimum resilience and transparency requirements, but their development is running significantly behind the pace of enterprise AI integration. Organizations that wait for regulatory clarity before developing concentration risk management practices are accepting a period of unmanaged exposure during which the operational, financial, and reputational consequences of concentration-driven failures will accumulate.

The path forward requires taking AI supply chain risk as seriously as enterprises now take cloud security risk — investing in visibility, designing for resilience, and applying to AI dependencies the same disciplined skepticism that Zero Trust principles brought to network perimeter assumptions. The market structure underlying enterprise AI is fragile in ways that are not visible in individual provider uptime metrics or security certifications. Making that fragility visible, and managing it systematically, is the work the security community needs to undertake now, before the next cascading failure makes the stakes undeniable.

References

[1] Carbon Credits. "NVIDIA Controls 92% of the GPU Market in 2025 and Reveals Next Gen AI Supercomputer." Carbon Credits, 2025.

[2] Silicon Analysts. "NVIDIA AI GPU Market Share 2026: ~80% of AI Accelerators." Silicon Analysts, 2026.

[3] Runtime News. "As AI Adoption Surges, AI Uptime Remains a Big Problem." Runtime News, 2026.

[4] Storyboard18. "Biggest AI Outages Since 2024: ChatGPT, Claude and Cloudflare Disruptions." Storyboard18, 2026.

[5] EU AI Act. "Article 51: Classification of General-Purpose AI Models as General-Purpose AI Models with Systemic Risk." European Parliament and Council, 2024.

[6] Federal Trade Commission. "FTC Focus: Enforcers Study AI Innovation and Entrenchment." Proskauer (FTC report summary), 2025.

[7] Norton Rose Fulbright. "The UK CMA's Review of AI Foundation Models." Norton Rose Fulbright, 2024.

[8] Gartner. "Gartner Says Worldwide AI Spending Will Total $1.5 Trillion in 2025." Gartner, September 2025.

[9] Bloomsbury Intelligence and Security Institute. "Compute Concentration and Systemic Risk in the Digital Economy." BISI, 2025.

[10] Introl. "Custom Silicon Inflection 2026: Hyperscaler ASICs vs NVIDIA GPU." Introl, 2026.

[11] MindStudio. "Google Cloud vs AWS vs Azure Q1 2026 – Which Hyperscaler Is Winning the AI Infrastructure Race?." MindStudio, 2026.

[12] GMI Cloud. "GPU Cloud Cost Comparison: An AI Startup's Guide for 2025." GMI Cloud, 2026.

[13] ALM Corp. "ChatGPT Down: Complete 2025 Outage Guide." ALM Corp, 2025.

[14] Trending Topics EU. "Claude Outages Surge as Anthropic Chases 2026 Revenue Lead Over OpenAI." Trending Topics EU, 2026.

[15] AI Business. "AWS Outage Takes Down AI Applications, Many Others." AI Business, 2025.

[16] PYMNTS. "Anthropic Outage Shows AI Is Straining the Digital Stack." PYMNTS, 2026.

[17] IAPP. "The Hidden Fragility of AI Supply Chains: Why Traditional Risk Management Falls Short." IAPP, 2025.

[18] TrustArc. "AI Supply Chain Risk: The New Vendor Due Diligence." TrustArc, 2025.

[19] CryptoRank. "SEC's Gary Gensler Raises Alarm on Financial Stability Due to AI Monoculture." CryptoRank, 2024.

[20] Preprints. "Model Monoculture Risk: Systemic AI Convergence in Banking and Financial Markets." Preprints.org, 2026.

[21] EU AI Act. "Article 55: Obligations for Providers of General-Purpose AI Models with Systemic Risk." European Parliament and Council, 2024.

[22] Stanford CRFM. "Foundation Models under the EU AI Act." Stanford CRFM, 2024.

[23] National Law Review. "AI and Antitrust 2025: DOJ, FTC Scrutiny on Pricing & Algorithms." National Law Review, 2025.

[24] PYMNTS. "Capitol Hill Confronts AI's Growing Grip on Financial Services." PYMNTS, 2025.

[25] Congressional Research Service. "Artificial Intelligence and Derivatives Markets: Policy Issues." CRS, 2025.

[26] U.S. Government Accountability Office. "Artificial Intelligence: Use and Oversight in Financial Services." GAO-25-107197, May 2025.

[27] Information Technology Intelligence Consulting. "ITIC 2024 Hourly Cost of Downtime Report." ITIC, 2024.

[28] arXiv. "Artificial Intelligence and Systemic Risk: A Unified Model of Performative Prediction, Algorithmic Herding, and Cognitive Dependency in Financial Markets." arXiv:2604.03272, 2026.