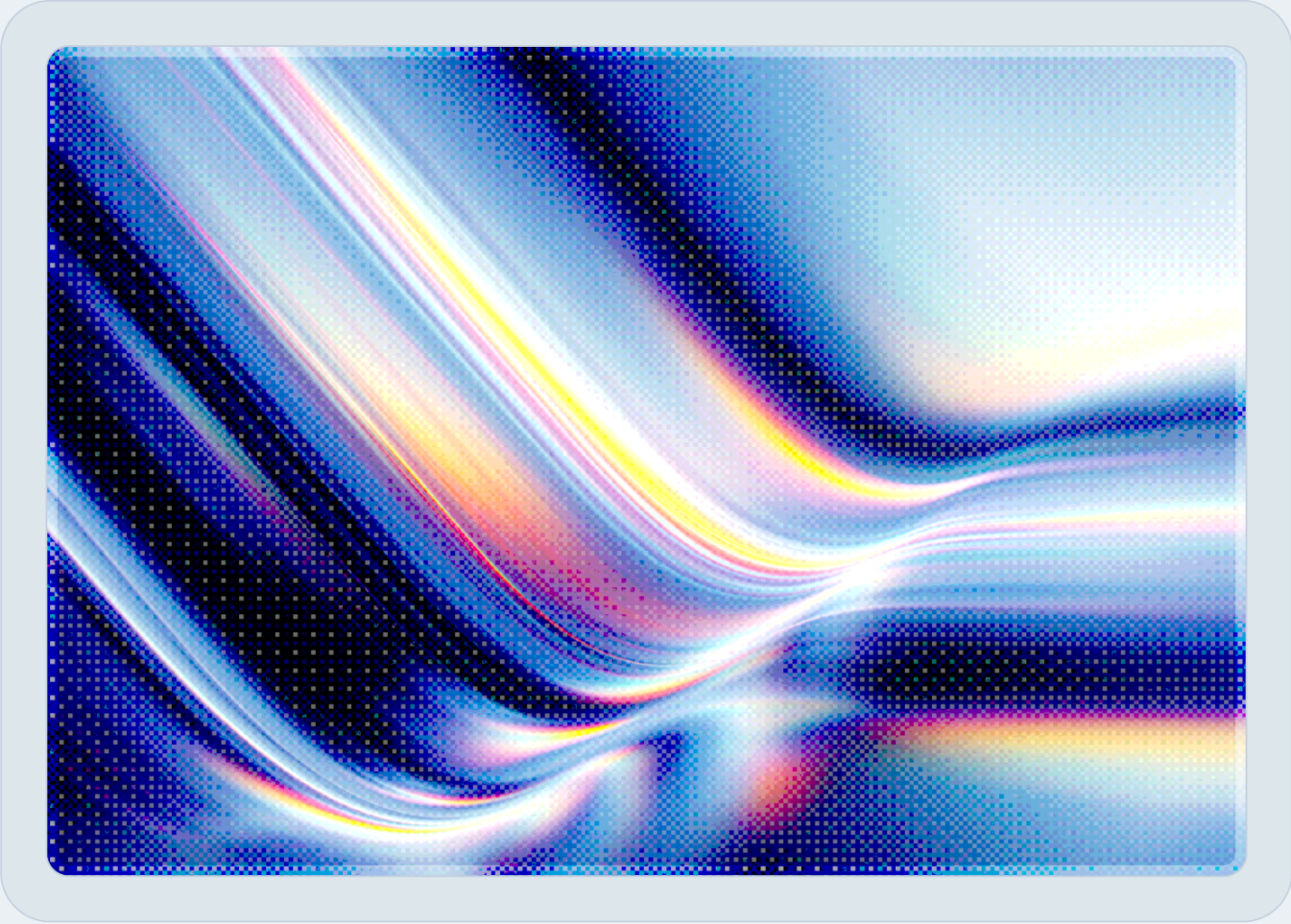
CSAI Foundation | Cloud Security Alliance

AI Development Stack Concentration Risk

When ML Frameworks Become Critical Infrastructure

2026-05-03

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- The global AI development stack exhibits significant market concentration across hardware, cloud compute, model distribution, and framework layers: a single firm holds more than 80% market share in three distinct AI infrastructure segments, and the top three firms collectively control more than 60% in three others, according to a 2025 OECD analysis of competition in AI infrastructure [1]. This is not a future risk — it is a present structural condition with active exploitation potential.
- NVIDIA controls approximately 92% of the discrete GPU market and a substantial share of AI accelerator revenue [2], while High Bandwidth Memory (HBM) — the specialized memory that makes GPU-scale AI practical — is sourced from just three suppliers: SK Hynix (~50%), Samsung (~40%), and Micron (~10%), with all 2025–2026 production capacity already committed [3].
- PyTorch dominates AI research workflows, with independent framework tracking analyses consistently placing its adoption above 70% of published papers [23], and the dependency ecosystems beneath PyTorch and other major AI frameworks have been directly targeted by attackers — most notably a December 2022 dependency confusion attack against PyTorch nightly that exfiltrated SSH keys, environment variables, and cloud credentials from affected developer machines [4].
- Hugging Face, which hosts over 1.41 million model repositories and functions as the de facto global distribution hub for AI models, suffered a breach of its Spaces platform in May 2024 that exposed authentication secrets and API tokens [5]; a subsequent analysis of the platform found over 352,000 suspicious issues across approximately 51,700 models [6], while JFrog's 2025 Software Supply Chain Report documented a 6.5× increase in malicious model uploads year-over-year through 2024 [7].
- The U.S. Department of Defense, CISA, NIST, and the OECD have each published guidance or framework material in the 14 months through May 2026 specifically addressing AI supply chain concentration and critical infrastructure risk [1][8][9][10], signaling a regulatory environment that will increasingly require organizations to document and mitigate these dependencies.

Background

The modern AI development stack is not a loosely coupled collection of interchangeable tools. It is a tightly integrated, hierarchically dependent architecture in which a small number of dominant platforms at each layer — hardware, cloud compute, model distribution, and software framework — provide services that have become practically irreplaceable for the organizations that rely on them. This architecture has followed a consolidation trajectory similar to, though arguably faster than, the broader cloud infrastructure market in its formative years, and the security implications are substantially higher because the systems now being built on this infrastructure are entering operational roles in critical sectors: power grids, financial systems, healthcare diagnostics, and autonomous transportation.

Concentration risk in technology infrastructure is not inherently catastrophic, but it does change the threat landscape in predictable ways. When a framework is used in more than 70% of AI research implementations [23] and a significant fraction of production deployments, a single vulnerability in that component becomes a vulnerability at scale — affecting thousands of organizations simultaneously rather than one at a time. Attackers understand this arithmetic. The December 2022 PyTorch supply chain attack exploited the assumption that a package used in PyTorch's nightly build pipeline was available only from a trusted internal index; the malicious `torchtriton` package uploaded to the public PyPI registry was downloaded 2,717 times before detection, silently exfiltrating credentials from developer machines [4]. The attack did not require a zero-day exploit. It required only an understanding of how the ecosystem's trust model worked and a gap that the dominant framework's maintainers had not yet closed.

This dynamic — where the very ubiquity of a platform creates a multiplier effect for attacks that successfully compromise it — defines the central security challenge of AI stack concentration. Organizations that have not mapped their exposure to these chokepoints are likely operating with incomplete threat models — AI supply chain dependencies represent an attack surface that perimeter and endpoint controls, however well-configured, are not designed to address.

Security Analysis

Hardware and Compute Concentration

The physical substrate of AI workloads — GPUs and the specialized memory that powers them — is controlled by a remarkably small number of organizations. NVIDIA's 92% share of the discrete GPU market means that a critical vulnerability in the CUDA runtime, driver stack, or associated tooling has potential impact across virtually every AI deployment [2]. This is not theoretical: NVIDIA issued security bulletins in September and October 2025 covering multiple CVEs in its CUDA Toolkit and display driver components, including race conditions, out-of-bounds reads, and uncontrolled DLL loading vulnerabilities [11][12]. These vulnerabilities span production inference infrastructure, model training pipelines, and individual developer workstations — with the specific risk profile varying by deployment context, but no segment categorically exempt from exposure.

The HBM supply bottleneck adds a different class of risk. The three HBM suppliers — SK Hynix, Samsung, and Micron — operate fabrication facilities that cannot be rapidly scaled; new capacity requires multi-year construction timelines. All production committed through 2026 has been allocated, and memory prices rose approximately 30% in late 2025 with further increases projected [3]. For organizations building or expanding AI infrastructure, this concentration means that supply chain disruptions, geopolitical events affecting South Korean or Taiwanese manufacturing, or security incidents at any of these suppliers have direct operational consequences. The January 2025 AI Diffusion Rule demonstrated that the policy environment around this layer of the stack is active and capable of rapid change; that rule was subsequently rescinded by the Trump administration in May 2025 — a sequence that reinforces, rather than undercuts, the observation that policy shifts in this domain can be abrupt and consequential for supply chain planning [13].

TSMC's role in advanced packaging — specifically CoWoS-L interposer technology required for high-density GPU assembly — introduces a further chokepoint. NVIDIA is reported to have secured more than 70% of TSMC's CoWoS-L capacity [3], which simultaneously limits competitors' ability to offer alternatives and concentrates the geopolitical risk associated with Taiwan's semiconductor ecosystem into a single supplier relationship.

Model Distribution as a Chokepoint

Hugging Face occupies a position in the AI ecosystem that has no close prior analog: a single privately held company functions as the primary distribution mechanism for AI models used by researchers and practitioners globally. With 1.41 million repositories and over 4.47 million unique model versions indexed

as of early 2025 [6], it functions more like a software package registry than a website — and it should be evaluated with the same security scrutiny applied to npm, PyPI, or Maven Central.

The May 2024 breach of the Hugging Face Spaces platform demonstrated that this analogy carries real risk. Attackers accessed authentication secrets and API tokens associated with organizations using the platform, a class of compromise that can propagate downstream to any application or pipeline that consumed models distributed through the affected accounts [5]. Hugging Face responded by implementing key management service (KMS) infrastructure and fine-grained token restrictions [24], but the incident illustrates the platform's fundamental attack surface: any organization that pulls a model from Hugging Face at deployment time has an implicit dependency on the security posture of both Hugging Face itself and the model's original uploader.

The scale of malicious content on the platform is significant. JFrog's 2025 Software Supply Chain Report documented a 6.5× year-over-year increase in malicious model uploads [7], while Protect AI's scanning partnership with Hugging Face identified over 352,000 suspicious issues across more than 51,700 models in the platform's public corpus [6]. Cisco's Foundation AI initiative has since deployed malware scanning at upload time for every public file submitted to Hugging Face [14], but this represents a recent mitigation applied to a corpus that already contains, by published estimates, over 352,000 flagged issues across more than 51,700 models.

Framework Ecosystem Vulnerabilities

The Python packaging ecosystem beneath PyTorch and TensorFlow has become a persistent attack surface as AI development has grown. AI projects typically carry substantially deeper dependency graphs than non-AI software, each additional package representing a potential entry point for supply chain compromise. This exposure is not theoretical: security researchers have consistently documented year-over-year increases in malicious packages targeting machine learning workflows, with attackers exploiting developer trust in familiar package names and the relative immaturity of AI toolchain security practices compared to broader software development.

The Ultralytics/YOLO incident in December 2024 is instructive as a template for this attack class. Attackers exploited a GitHub Actions script injection vulnerability to steal the maintainer's PyPI publish token, then injected cryptocurrency mining code into four consecutive releases of the Ultralytics package — a widely-used computer vision library — before the compromise was detected [16]. The attack chain did not target the package's code directly; it targeted the CI/CD pipeline that produced the package, exploiting the implicit trust that end-users extend to packages published by known maintainers [15]. The prevalence of GitHub Actions-based CI/CD in AI repositories makes this attack pattern broadly applicable: security researchers have documented systematic misconfiguration patterns in AI and ML

repository workflows — including unpinned action references, overly permissive tokens, and script injection pathways of the type exploited in the Ultralytics incident — that leave many AI development pipelines vulnerable to an analogous compromise [15][17].

TensorFlow's CVE history warrants attention from organizations that treat framework maturity as a proxy for security. CVE-2024-3660, a Keras arbitrary code injection vulnerability with a CVSS score of 9.8, demonstrated that deserialization vulnerabilities in model loading code — a pathway used routinely in production deployment pipelines — can yield full code execution [18]. The LangChain-core serialization injection vulnerability (CVE-2025-68664) followed a similar pattern, affecting a widely-deployed orchestration layer that sits atop multiple AI frameworks [25]. Both cases illustrate that the attack surface for AI systems is not limited to the model or data layers; the orchestration and serialization code that loads, transforms, and serves AI outputs represents a distinct and often under-scrutinized attack surface.

Cloud Infrastructure Dependencies

The three largest cloud providers — AWS (approximately 33% market share), Azure (approximately 23%), and Google Cloud (approximately 12%) — collectively control roughly two-thirds of global cloud infrastructure revenue [19]. For AI workloads specifically, this concentration is more pronounced: all three providers have made substantial proprietary investments in AI-optimized hardware (AWS Trainium, Google TPUs, Azure's partnership with NVIDIA for custom infrastructure) that creates additional lock-in beyond standard cloud portability concerns.

Enterprise multi-cloud adoption has increased substantially in recent years as organizations seek to diversify compute dependencies [20]. However, multi-cloud strategies introduce their own complexity: security postures, identity models, and network architectures must be maintained consistently across providers, and AI-specific services — managed training pipelines, inference endpoints, vector databases — typically have no direct functional equivalent across providers, limiting true portability. In the authors' assessment, organizations that have distributed their AI workloads across multiple clouds for availability reasons may still face significant concentration risk in the AI-specific service layer, where equivalent managed services across providers are rarely interchangeable.

Recommendations

Immediate Actions

The most immediately actionable control is a dependency inventory: for every production AI system, organizations should document the GPU vendor, cloud provider, model distribution source, ML framework, and primary Python dependencies. This inventory is the prerequisite for any meaningful concentration risk assessment and should be treated with the same operational urgency as an asset inventory for traditional software infrastructure. Without it, organizations cannot determine which of the concentration risks described in this note apply to their specific deployment environment.

Model provenance verification should accompany any model pulled from Hugging Face or other external repositories. Implementing hash verification for externally sourced models, combined with pinning model versions in deployment manifests rather than pulling from a floating "latest" tag, reduces exposure to both malicious model uploads and accidental consumption of compromised model versions. A review of the publishing account's history and reputation is warranted before any externally hosted model enters a production context; high download counts and star ratings are not substitutes for security evaluation.

Organizations running AI development workflows should audit GitHub Actions configurations and other CI/CD pipeline settings for misconfigurations documented in recent attack campaigns: unpinned action references, overly permissive tokens, and script injection pathways. The Ultralytics attack pattern is well-documented and repeatable, and its core exploit — implicit trust extended to the pipeline that produces a package rather than the package itself — is present in many AI repositories where developers focus security attention on the model and data layers while leaving pipeline infrastructure underexamined.

For organizations running NVIDIA GPUs in production, CUDA driver updates should be treated with the same urgency as OS-level security patches. The September and October 2025 bulletins addressed vulnerabilities with active exploitation potential in AI infrastructure contexts [11][12]; a delayed patch cycle for GPU driver components is no less risky than delayed patching for other system-level software.

Short-Term Mitigations

At the dependency level, adopting lockfiles and explicit version pinning throughout AI development workflows directly addresses the dependency confusion attack class demonstrated against PyTorch. Using `poetry.lock`, `requirements.txt` with explicit version constraints, or equivalent mechanisms prevents silent dependency substitution. Some dependency management tools include

delayed-ingestion features that avoid pulling packages uploaded within a defined recency window, providing an additional layer of defense against packages that exploit the brief interval between upload and community review.

The Open Neural Network Exchange (ONNX) format warrants evaluation as an intermediate representation for production model workflows [21]. ONNX enables models trained in PyTorch or TensorFlow to be exported to a common interchange format and executed via ONNX Runtime across different hardware backends — a meaningful hedge against framework-specific vulnerabilities and a partial mitigation for framework lock-in. ONNX adoption carries practical constraints that organizations should understand before committing to it as a portability strategy: not all PyTorch operators export cleanly to the ONNX specification, quantization behavior can differ between frameworks, and ONNX Runtime performance varies by hardware backend. Organizations should evaluate ONNX as a portability option where its operator support is sufficient for the specific model architecture in question, rather than treating framework portability as a guaranteed outcome of ONNX adoption.

Organizations with AI systems in operational roles should establish and test contingency capacity on a secondary cloud provider. The goal is not continuous multi-cloud operation — which adds significant complexity to security posture, identity management, and network architecture — but the verified ability to shift inference workloads under degraded primary conditions. This capability should be tested under realistic conditions, not merely designed; an untested failover path is not a mitigation.

Strategic Considerations

AI infrastructure concentration is a structural market condition, not a configuration error, and it will not resolve quickly. The OECD's finding that in some AI infrastructure segments a single firm holds more than 80% market share [1] reflects economic dynamics — network effects, capital requirements, and talent concentration — that take years to shift. Organizations should plan accordingly, treating concentration risk as a durable feature of the AI infrastructure landscape rather than a transitional state that will self-correct.

From a governance perspective, the regulatory direction is consistent: NIST's April 2026 concept note for a Trustworthy AI in Critical Infrastructure profile [9], CISA's May 2025 AI data security guidance [8], and the EU AI Act's classification of AI systems used in critical infrastructure as high-risk [22] all signal an environment in which formal concentration risk governance is likely to become expected, if not required. Organizations that build this capability now — through asset inventories, concentration risk frameworks, and supplier diversity roadmaps — will be better positioned than those who address it reactively under regulatory pressure.

The DoD's March 2026 guidance on AI/ML supply chain risks and mitigations recommends extending existing supply chain risk management (SCRM) programs to explicitly cover model supply chains, data pipelines, and AI framework dependencies, with contractual requirements for vendor transparency including software bills of materials (SBOMs) [10]. Security teams that have invested in software supply chain security programs for traditional software should evaluate whether those programs extend meaningfully to the AI-specific layers of the stack described in this note.

CSA Resource Alignment

The risks described in this research note map directly to several CSA frameworks that provide actionable control guidance.

The **AI Controls Matrix (AICM) v1.0** addresses AI supply chain security as a distinct control domain, covering shared responsibility models for model providers, orchestrated service providers, and AI customers. The AICM's supply chain controls are the most directly applicable framework for organizations conducting a formal gap assessment against the concentration risks described here. In particular, the AICM guidance for orchestrated service providers (OSPs) addresses the scenario where a single model hub or framework becomes a dependency for multiple downstream applications. Organizations should treat the AICM as the primary reference framework for AI supply chain governance, noting that it is a superset of the Cloud Controls Matrix (CCM) and replaces CCM as the default reference for AI-specific workloads.

MAESTRO (CSA's Agentic AI Threat Modeling framework) provides relevance in the context of agentic pipelines that consume models or packages at runtime. An agentic AI system that pulls a model from Hugging Face, installs a Python dependency via pip, or calls an inference endpoint at runtime inherits the supply chain risk associated with each of those operations. MAESTRO threat modeling for agentic systems should explicitly include the dependency and model acquisition steps as threat surfaces, not only the agent's behavior once deployed.

The **STAR (Security Trust Assurance and Risk)** registry program is directly applicable to cloud providers and ML framework vendors as third-party AI infrastructure suppliers. Organizations seeking evidence of security controls at AWS, Azure, GCP, or major ML platform vendors should evaluate STAR certifications as part of their supplier due diligence process, supplemented by AI-specific contract provisions that require SBOM disclosure and notification of material supply chain incidents.

CSA's **Zero Trust guidance** is relevant to the model distribution risk: a zero trust posture applied to AI model consumption treats every externally sourced model as untrusted until verified, regardless of the reputation of the hosting platform. This contrasts with the prevalent practice of implicitly trusting models hosted on Hugging Face or similar platforms based on download count or star rating.

References

[1] OECD. "Competition in Artificial Intelligence Infrastructure." OECD Publishing, 2025.

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

[3] BCD Video. "Inside the 2025–2027 Compute Crunch: What Supply Chain Volatility Really Means for You." BCD Video, 2025.

[4] PyTorch. "Compromised PyTorch-nightly Dependency Chain Between December 25th and December 30th, 2022." PyTorch Blog, January 2023.

[5] OECD AI Incident Monitor. "Hugging Face Spaces Platform Breach Exposes Authentication Secrets." OECD.ai, May 2024.

[6] Hugging Face / Protect AI. "4M Models Scanned: Protect AI + Hugging Face 6 Months In." Hugging Face Blog, 2025.

[7] JFrog Security Research. "2025 Software Supply Chain State of Security Research Report." JFrog, April 2025.

[8] CISA. "CISA Releases Guidance on AI Data Security." CISA.gov, May 2025.

[9] Industrial Cyber. "NIST Develops Trustworthy AI in Critical Infrastructure Profile to Align Risk, Resilience, and Infrastructure Security." Industrial Cyber, April 2026.

[10] U.S. Department of Defense. "Artificial Intelligence and Machine Learning: Supply Chain Risks and Mitigations." DoD, March 2026.

[11] NVIDIA. "Security Bulletin: NVIDIA CUDA Toolkit – September 2025." NVIDIA Product Security, September 2025.

[12] NIST. "CVE-2025-23280: NVIDIA Display Driver Use-After-Free Vulnerability." National Vulnerability Database, October 2025.

[13] U.S. Department of Commerce, Bureau of Industry and Security. "Framework for Artificial Intelligence Diffusion." Federal Register, January 2025. (Subsequently rescinded May 2025.)

[14] Cisco Security. "Cisco's Foundation AI Advances AI Supply Chain Security With Hugging Face." Cisco Blog, 2025.

[15] PyPI Blog. "Supply-Chain Attack Analysis: Ultralytics." Python Package Index Blog, December 2024.

[16] Socket. "PyPI on Ultralytics Breach: No Security Flaws in PyPI Exploited." Socket.dev, December 2024.

[17] TrueFoundry. "Supply Chain Attacks in AI: What the LiteLLM Incident Reveals." TrueFoundry Blog, 2025.

[18] Oligo Security. "TensorFlow Keras Downgrade Attack: CVE-2024-3660." Oligo Security Blog, 2024.

[19] Business Stats. "Cloud Market Share 2026: AWS vs Azure vs Google Revenue and Full Stats." BusinessStats, 2026.

[20] CrispIdea. "Cloud Computing in 2026: AI, Edge & Multi-Cloud Trends." CrispIdea, 2026.

[21] ONNX. "ONNX: Open Neural Network Exchange." ONNX Project, 2025.

[22] European Commission. "Regulatory Framework for Artificial Intelligence (EU AI Act)." European Commission, 2024.

[23] Papers With Code. "Machine Learning Framework Trends." Papers With Code, 2025.

[24] Hugging Face. "Sharing Space Security Incidents and Fixes." Hugging Face Blog, May 2024.

[25] NIST. "CVE-2025-68664: LangChain-Core Serialization Injection Vulnerability." National Vulnerability Database, 2025.