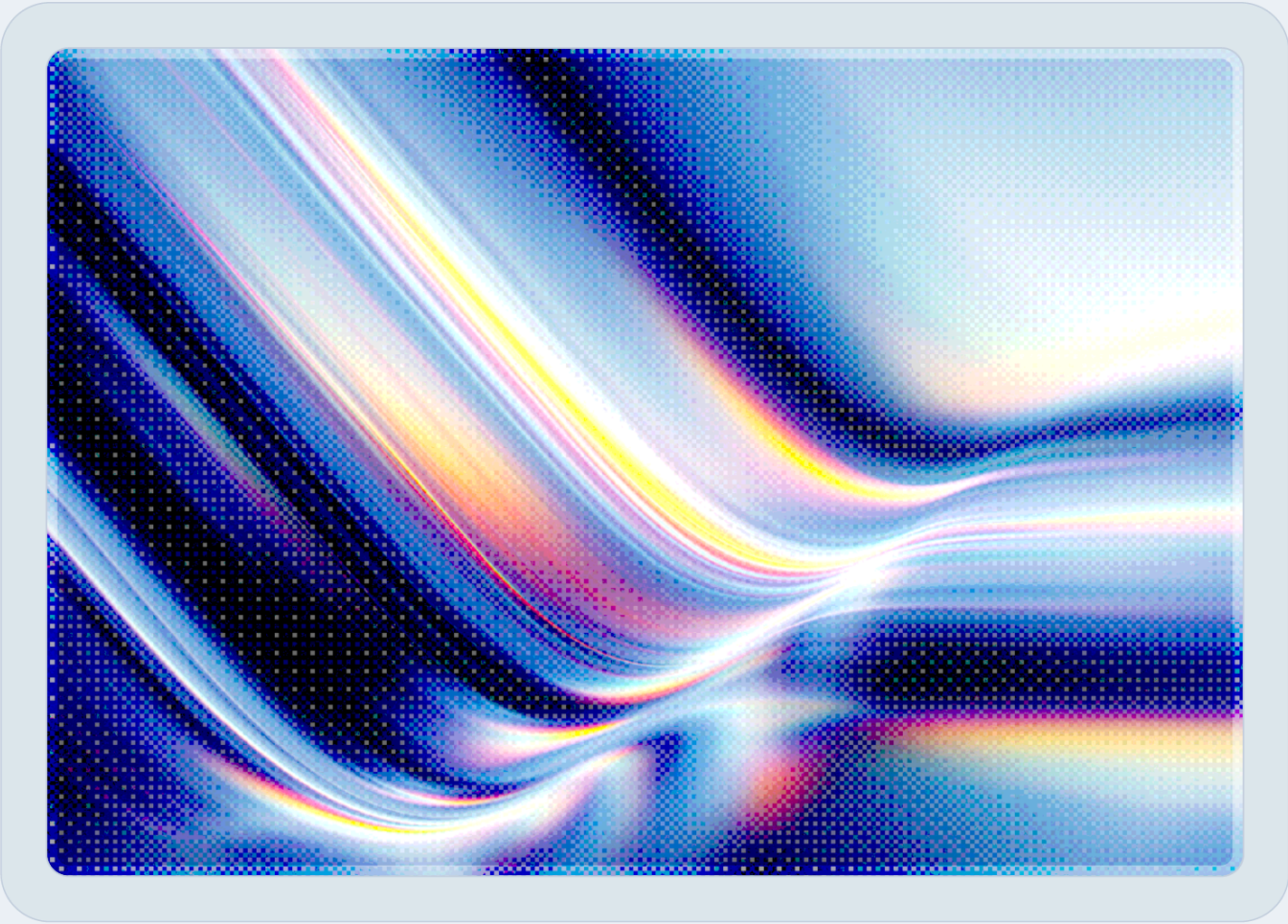
CSAI Foundation | Cloud Security Alliance

Frontier AI as Geopolitical Lever

Export Controls, Sovereign AI Risk, and Enterprise Dependency

2026-06-21

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Table of Contents

Executive Summary 4

Introduction: AI at the Intersection of Technology and Geopolitics 5

The Export Control Landscape: Chips as Instruments of State Power 6

The Sovereign AI Race: Nations Reclaiming Compute Independence 8

China's Strategic Response: Building a Parallel AI Stack 9

Enterprise Dependency: The Hidden Third-Party Risk 11

Geopolitical Risk Scenarios and Their Enterprise Implications 12

A Framework for Enterprise AI Resilience 14

Conclusions and Recommendations 16

CSA Resource Alignment 17

References 19

Executive Summary

Frontier artificial intelligence has crossed a threshold that has fundamentally transformed how governments treat advanced technology: it has become a tool of statecraft comparable in strategic weight to energy resources, financial systems, and military capability. Governments are no longer treating advanced AI as a technology sector subject to ordinary trade rules. Instead, they are treating frontier compute capacity, AI model access, and semiconductor supply chains as sovereign assets requiring direct state management. For enterprises, this shift has created a new category of risk that sits at the intersection of technology procurement, geopolitical exposure, and operational continuity.

The immediate policy landscape is defined by export controls on advanced semiconductors, which have oscillated between systematic multilateral frameworks and transactional bilateral arrangements depending on the prevailing U.S. administration. The Biden administration's January 2025 AI Diffusion Rule established a three-tier system of controlled AI chip exports covering most of the world; the Trump administration rescinded that framework in May 2025 and replaced it with a more transactional approach, including revenue-sharing arrangements with semiconductor companies selling to China. Simultaneously, the Chip Security Act advanced through Congress in 2026 with bipartisan backing, proposing to embed location-verification technology directly into export-controlled chips—a measure that, if implemented, would reshape hardware supply chains worldwide.

National responses to this export-control environment have been large-scale and have accelerated significantly. Commercial estimates suggest sovereign wealth funds committed as much as $120 billion to AI infrastructure in 2025–2026, and South Korea launched coordinated sovereign AI programs spanning semiconductor fabrication, government research, and industrial transformation, with Samsung's $310 billion semiconductor investment commitment anchoring the effort [11]. France pledged €109 billion in AI investment in February 2025. The European Union adopted a sweeping European Technological Sovereignty Package in June 2026, responding to findings that over 80 percent of the EU's key digital products, services, and intellectual property originate outside the bloc. China, denied access to the most advanced U.S. chips, accelerated its sovereign AI stack: DeepSeek V4 runs on Huawei's Ascend 950PR processor with no dependencies on NVIDIA CUDA or American software, and Chinese domestic chips captured approximately 41 percent of the Chinese AI chip market in 2025—up from single digits before 2023.

For enterprise security leaders, these developments create a distinctive set of risks that existing third-party risk frameworks were not designed to address. An April 2026 Zapier survey found that 74 percent of enterprises would face significant disruption if they lost access to their primary AI vendor [23].

Anthropic's annual revenue run-rate surpassed $30 billion in 2026, with more than 1,000 business customers each spending over $1 million annually; OpenAI closed a $122 billion funding round and reported that enterprise revenue now exceeds 40 percent of total income. The concentration of production-grade AI capability in a handful of U.S.-headquartered providers means that regulatory action, sanctions, infrastructure failure, or geopolitical conflict affecting those providers can ripple directly into enterprise operations globally.

This whitepaper provides a structured analysis of these dynamics and a practical risk management framework for enterprise security leaders and CISOs.

Introduction: AI at the Intersection of Technology and Geopolitics

The history of strategically controlled technologies—fissile materials, cryptography, satellite components—offers a clear pattern: when a capability becomes sufficiently consequential, states assert control over its production, distribution, and use. Advanced AI has entered this category. The compute infrastructure required to train frontier AI models, the semiconductors that power inference at scale, and the models themselves are now subject to export licensing, investment screening, and legislative mandates in every major economy. Enterprises that depend on these systems for operational AI workloads are increasingly caught between competing national strategies.

The consequences of this shift are not primarily theoretical. Export control policy changes directly affect the availability and pricing of AI infrastructure services. Geopolitical tensions constrain where AI compute can be physically located and whether cross-border data flows supporting AI applications remain permissible. The policy oscillation visible in the United States between January 2025 and mid-2026—from a systematic three-tier export control framework to bilateral transactional deals to embedded-chip tracking legislation—illustrates how quickly the regulatory ground can shift beneath enterprise procurement decisions. Organizations that built AI infrastructure strategies around stable regulatory assumptions are discovering that those assumptions may not hold across a two-year planning horizon.

This paper addresses three interconnected analytical domains. The first is the export control landscape: how governments are using semiconductor and compute access as instruments of strategic competition, and what enterprises need to understand about how those controls work and change. The second is sovereign AI: the wave of national investment programs aimed at reducing dependency on foreign-controlled AI infrastructure, and what competitive dynamics those programs create. The third is

enterprise AI dependency: the concentrated market structure of frontier AI, how vendor lock-in mechanisms operate, and the risk scenarios that enterprises face as geopolitical pressures increase. The paper concludes with a practical framework for building AI resilience into enterprise security programs.

The World Economic Forum's Global Cybersecurity Outlook 2026, drawing on responses from 804 qualified participants across 92 countries, found that 66 percent of organizations modified their cybersecurity strategy due to geopolitical instability, and that geopolitics remains the top factor influencing overall cyber risk mitigation [1]. AI is central to this dynamic: 94 percent of respondents identified AI as the single most significant driver of cybersecurity change in 2026, and 87 percent flagged AI-related vulnerabilities as the fastest-growing cyber risk throughout 2025 [1]. These findings reflect an enterprise security environment in which AI is simultaneously a risk management tool and a source of new, geopolitically amplified vulnerability.

The Export Control Landscape: Chips as Instruments of State Power

The fundamental logic underlying AI export controls was that advanced AI capability is inseparable from advanced computing hardware—an assumption that subsequent events, particularly China's frontier AI development, would substantially complicate. Training frontier AI models requires clusters of specialized graphics processing units capable of performing billions of floating-point operations per second at the scale of thousands of nodes working in parallel. The primary producer of chips capable of this workload—NVIDIA—is a U.S. company, and its hardware requires a U.S.-origin software stack (the CUDA ecosystem) to operate at maximum efficiency. This geography of production gave the United States a lever it has deployed with increasing assertiveness since the October 2022 semiconductor export restrictions targeting China.

The most systematic attempt to extend this leverage globally was the Biden administration's Framework for Artificial Intelligence Diffusion, published in the Federal Register on January 15, 2025 [2]. The rule created a three-tier structure governing AI chip exports from the United States to the rest of the world. Tier 1, comprising the United States and eighteen close allies, faced essentially no restrictions. Tier 2, which encompassed most of the remaining world, could access chips through a data center Validated End User program operating under a presumption of license approval, but with the constraint that Tier 1-headquartered companies could not deploy more than 25 percent of their total AI compute capacity in Tier 2 jurisdictions, with no more than seven percent in any single Tier 2 country. Tier 3, including China,

Russia, and other embargoed destinations, remained subject to near-total restriction [3]. The framework was designed to preserve U.S. AI leadership while ensuring that allied nations could participate in the buildout of frontier AI infrastructure, albeit on terms controlled from Washington.

The rule never fully took effect. The Trump administration's Department of Commerce formally rescinded the AI Diffusion Framework on May 13, 2025—just two days before its 120-day delayed compliance period would have activated [4]. The rescission left the framework without a direct replacement; the administration instead pursued a series of transactional bilateral arrangements. In August 2025, NVIDIA was granted export licenses to sell its H20 chips to China on the condition that the company pay 15 percent of its revenues from those sales to the U.S. government [29]. The arrangement was extended in December 2025 to NVIDIA's more advanced H200 chips at a 25 percent revenue-sharing rate, with similar terms applied to AMD and Intel products [29].

On January 13, 2026, BIS issued a final rule formalizing a revised export licensing posture for commercially available NVIDIA H200 and AMD MI325X-equivalent chips destined for China, changing the review standard from a presumption of denial to a case-by-case review conditioned on specific technical, business, end-user, and U.S. market certifications [5][28]. The effect is a more permissive but considerably more uncertain regulatory environment: access to the most capable chips is now available to Chinese customers in principle, but contingent on administrative approval processes that can change with executive direction.

The practical instability of this approach has motivated bipartisan legislative action. The Chip Security Act, which advanced through the House Foreign Affairs Committee with bipartisan support and had cleared Congress by March 2026 [6], takes a structurally different approach: rather than relying on licensing decisions administered at the executive level, it proposes to embed location-verification technology directly into export-controlled chips. The legislation would require the Commerce Department to mandate continuous physical location reporting on advanced AI chips within 180 days of enactment, and to require companies to report any suspected diversions [6]. The Semiconductor Industry Association has opposed blanket on-chip mandates, citing concerns about implementation complexity and the risk of undermining global confidence in U.S. semiconductor products [7]. The fundamental intent of the legislation, however, reflects a recognition that administrative export controls are vulnerable to policy reversal and evasion alike, and that durable control requires hardware-level enforcement.

For enterprise security leaders, the practical consequences of this environment are less about legal compliance—most enterprises are not importing controlled chips themselves—and more about supply chain stability, pricing, and infrastructure availability. Cloud AI services and inference APIs are priced in part on the underlying compute costs of their providers, and those costs are affected by export control regimes that influence global chip availability and distribution. Enterprises with multinational operations

must also assess whether their AI compute arrangements, particularly third-party AI services used by subsidiaries in Tier 2 countries, are consistent with evolving export control requirements. The RAND Corporation's analysis of the AI Diffusion Framework concluded that export controls attempting to create a U.S.-led global AI ecosystem require sustained multilateral coordination to be effective—coordination that the transactional approach taken since May 2025 has not prioritized [8].

The Sovereign AI Race: Nations Reclaiming Compute Independence

The simultaneous escalation of U.S. AI export controls and the recognition that advanced AI is foundational infrastructure for economic competitiveness and national security has produced a global wave of sovereign AI investment without precedent in recent technology policy history. Commercial estimates suggest sovereign wealth funds committed as much as $120 billion to data centers, semiconductor fabrication plants, and high-performance computing networks in 2025 and 2026 alone [9]. Globally, spending on sovereign AI systems is projected to surpass $100 billion in 2026, reflecting both the scale of the perceived strategic gap and the urgency with which governments are moving to close it [10].

The individual country commitments are substantial. France announced €109 billion in total AI investments in February 2025, combining the France 2030 initiative with significant private and international contributions, explicitly framed as a response to both U.S. export control uncertainty and the need to develop AI systems not subject to foreign jurisdictional control [14]. South Korea launched coordinated sovereign AI programs spanning semiconductor fabrication, government research and development, industrial transformation, and infrastructure, with Samsung committing $310 billion in semiconductor investment and government programs providing substantial complementary capital [11]. Canada established a Sovereign AI Compute Strategy with dedicated funding for a nationally owned supercomputing system, and Japan committed significant public capital to a national AI infrastructure initiative.

The European Union's response has been the most institutionally elaborate. The European Commission adopted the European Technological Sovereignty Package on June 3, 2026, comprising measures to strengthen EU capacity in semiconductors, artificial intelligence, cloud, and open source software [12]. The package follows findings that the EU currently relies on non-EU countries for over 80 percent of key digital products, services, infrastructure, and intellectual property [13]. The Commission has estimated the investment requirement at approximately €120 billion for semiconductors, €200 billion for data centers by 2036, €100 billion for cloud and AI, and €2 billion for open-source software over seven years

[12]. The EU Cloud and AI Development Act specifically targets reduction of strategic dependencies by creating better conditions for businesses, researchers, and public administrations to deploy AI on European infrastructure [14]. These efforts built on a France-Germany Summit on European Digital Sovereignty held in November 2025, which identified AI, data, and public infrastructure as priority areas and launched a joint task force.

The strategic logic behind these investments extends beyond economic competition. Governments are increasingly treating AI compute capacity as critical national infrastructure on par with power grids and communication networks. The International Institute for Strategic Studies characterized the sovereign AI movement as representing "pathways to strategic autonomy"—the ability to operate essential AI-dependent functions without relying on infrastructure subject to foreign regulatory control, sanctions risk, or policy change [15]. The concern is well-founded: an organization that depends on a foreign-controlled AI service for essential functions faces the possibility that geopolitical developments affecting that service's home country—sanctions, conflict, regulatory change, or provider decision to exit a market—could disrupt operations with little warning. Governments are making large capital commitments to reduce that exposure; enterprises face the same underlying risk without comparable state resources to hedge it.

The McKinsey Global Institute's analysis of sovereign AI ecosystem requirements found that a credible sovereign AI capability requires not just raw compute but a deep in-country infrastructure including data centers, high-density GPU clusters, cloud platforms, subsea connectivity, reliable energy supply, and a skilled technical workforce [16]. This is not a capability that most governments can assemble quickly, nor can most enterprises. The gap between current sovereign AI ambition and deployable sovereign AI capability is likely to persist for several years barring further unexpected capability breakthroughs of the kind demonstrated by DeepSeek, during which time the underlying dependency on U.S.-controlled frontier AI infrastructure is expected to remain substantially unchanged for most of the world outside China.

China's Strategic Response: Building a Parallel AI Stack

China's trajectory in frontier AI represents the most extensively documented and strategically significant near-term example of sovereign AI strategy in action, precisely because it has been driven by the most stringent export control restrictions. The October 2022 export controls and their subsequent tightening

severely constrained Chinese AI developers' access to NVIDIA's most advanced training hardware. The U.S. government's expectation was that this constraint would materially slow China's progress toward frontier AI capability. That expectation has been substantially revised.

The release of DeepSeek R1 in January 2025 was the most visible evidence of the gap between the assumed effectiveness of chip export controls and the actual state of Chinese AI research [17]. The model achieved performance approaching leading American frontier systems on standard benchmarks, reportedly developed at a fraction of the compute cost of comparable American models and without access to restricted chips. The European Union Institute for Security Studies characterized DeepSeek R1 as dealing "a serious blow to the technical and business model long championed by US tech giants," noting that the system's architectural efficiency demonstrated that frontier AI performance is not solely a function of raw compute availability [17]. The Council on Foreign Relations described DeepSeek V4, released subsequently, as "signals a new phase in the U.S.-China AI rivalry" representing a qualitative shift in the competitive dynamic [18].

The more strategically significant development is the construction of a complete, vertically integrated AI infrastructure stack with no American software dependencies. DeepSeek V4 was designed to run on Huawei's Ascend 950PR processor rather than NVIDIA or AMD silicon, and the engineering work invested in migrating the system from NVIDIA's CUDA ecosystem to Huawei's proprietary CANN framework produced a fully functional end-to-end AI stack spanning chip architecture, training frameworks, and inference software [19]. The U.S.-China Economic and Security Review Commission concluded in a March 2026 analysis that China's strategy of open-sourcing AI models while building sovereign hardware and software infrastructure serves a dual purpose: it frames China as a global provider of open AI development while propelling Chinese models toward becoming industry standards in markets not aligned with the U.S. technology ecosystem [20][27].

The market effects of this strategy are measurable. Chinese domestic chips captured approximately 41 percent of the Chinese AI chip market in 2025—up from a market share below 10 percent before 2023—with approximately half of domestic sales coming from Huawei [19]. In the global model ecosystem, Chinese fine-tuned or derivative models constituted 63 percent of new fine-tuned or derivative releases on Hugging Face in September 2025, indicating that Chinese open-source AI activity has achieved a scale of output that shapes the global model landscape independent of export control restrictions [20]. The CSIS analysis of DeepSeek, Huawei, and export controls characterized the competitive implications as "a new phase" in which the assumption that hardware restrictions translate directly into capability restrictions must be substantially qualified [21].

For enterprises with operations in regions where Chinese AI providers are dominant—including significant portions of Southeast Asia, the Middle East, and Africa—this parallel AI stack creates distinct considerations. Chinese AI platforms operating on sovereign hardware and software infrastructure may

be subject to different regulatory requirements than Western-controlled systems, including data localization obligations, government access to training data or model outputs, and alignment requirements that differ from those applied in Western regulatory environments. The existence of two AI ecosystems—one centered on U.S. providers and hardware, one increasingly centered on Chinese-origin infrastructure—creates procurement decisions with geopolitical implications that most enterprises have not yet incorporated into their vendor risk frameworks.

Enterprise Dependency: The Hidden Third-Party Risk

The enterprise AI market has concentrated rapidly around a small number of frontier model providers. Anthropic's annual revenue run-rate surpassed $30 billion in 2026, with more than 1,000 enterprise customers each spending in excess of $1 million per year [22]. OpenAI closed a $122 billion funding round and reported that enterprise revenue exceeds 40 percent of total income, with a clear strategic shift toward large enterprise contracts and services [22]. Google's Gemini family and Microsoft Azure AI services serve additional large segments of the enterprise market. The combined effect is that production AI workloads for the majority of large enterprises worldwide depend on a handful of providers, each headquartered in the United States, each subject to the full range of U.S. regulatory and geopolitical risk.

An April 2026 Zapier survey found that 74 percent of enterprises would face significant disruption if they lost access to their primary AI vendor [23]. This figure reflects how rapidly AI has moved from experimental deployment to operational integration. Where the first generation of enterprise AI adoption in 2023–2024 was characterized by isolated pilots and productivity tools, the 2025–2026 generation has embedded AI into customer-facing services, internal knowledge management, code generation, decision support, and automated workflows. Each of these integrations creates a dependency that is costly to reverse: switching AI providers in a mature deployment requires rewriting applications, migrating fine-tuned model weights, rebuilding integrations, retraining prompt libraries, and renegotiating data agreements. Unlike infrastructure commodities where substitution is relatively straightforward, AI dependencies are often deeply intertwined with proprietary data, customization work, and operational processes.

The lock-in mechanisms operating in the enterprise AI market differ from those that characterized earlier generations of technology vendor dependency. API format incompatibility is one dimension: while standardization efforts around common model interfaces exist, production deployments frequently rely on provider-specific features, context window specifications, tool-calling protocols, and output formatting that do not transfer directly between providers. Model behavior is another dimension: a fine-tuned model trained on proprietary enterprise data using one provider's infrastructure produces outputs

calibrated to that provider's base model; migrating to a different base model changes the output characteristics and may require revalidation of downstream applications. These technical switching costs compound commercial switching costs as providers move toward multi-year enterprise contracts with usage commitments, preferred pricing tied to volume, and professional services relationships.

The structure of provider relationships is also evolving in ways that increase dependency risk. Both Anthropic and OpenAI have expanded into enterprise services, moving beyond pure API access toward AI application development, agent deployment, and managed infrastructure [24]. This vertical integration strategy means that the same provider supplying the underlying model is also increasingly involved in the application layer built on top of it, deepening the operational entanglement and creating competitive dynamics that deserve attention in vendor risk assessments. A foreseeable consequence of this vertical integration is that an enterprise that has built customer-facing applications on a frontier AI provider's infrastructure may find itself competing with that same provider for end customers if the provider enters adjacent markets.

The concentration of AI compute in large data center facilities creates a secondary dependency risk at the infrastructure level. Disrupting a major AI provider's data center infrastructure—whether through cyberattack, natural disaster, power failure, or cascading technical fault—would affect all enterprises relying on that provider simultaneously. HSToday's analysis of concentration risk in digital ecosystems noted that "taking out the centres supporting major AI providers like Google and OpenAI would be just as disruptive as hitting traditional critical national infrastructure like a power grid," a comparison that illustrates the scale of systemic exposure now embedded in enterprise AI dependency [25]. The 2025 WEF Global Cybersecurity Outlook report documented that cascading failures linked to cloud service concentration have already demonstrated how localized incidents generate cross-sector consequences affecting healthcare, logistics, financial services, and manufacturing.

Geopolitical Risk Scenarios and Their Enterprise Implications

Understanding the abstract risks created by export controls and market concentration requires grounding them in concrete scenarios that enterprise security leaders can assess and plan against. The following scenarios are illustrative rather than exhaustive, drawn from current geopolitical dynamics and the structural dependencies analyzed in preceding sections.

The scenario most directly shaped by current policy volatility is policy-driven service restriction, in which regulatory action in the United States or a host country changes the terms of AI service availability for enterprises operating in specific jurisdictions. The rescission of the AI Diffusion Rule and its replacement with transactional bilateral arrangements created a period in which AI compute arrangements for enterprises with offshore data center operations—particularly in Tier 2 countries—were uncertain with respect to compliance. A future tightening of controls, renewed multilateral framework, or regulatory action targeting specific AI use cases (such as AI used in defense, critical infrastructure, or sensitive sectors) could restrict access to AI services with limited advance notice. Enterprises with operations in jurisdictions that become the subject of sanctions or export restrictions—as has occurred for Russia-based operations since 2022—could face loss of AI service access as providers comply with regulatory requirements, potentially within days.

A second scenario involves infrastructure concentration failure, in which a significant cyberattack, supply chain incident, or operational failure affecting a major AI provider disrupts services across all enterprises simultaneously. The WEF Global Cybersecurity Outlook 2026 found that 64 percent of organizations are now accounting for geopolitically motivated cyberattacks targeting critical infrastructure [1], and national-level AI infrastructure is an increasingly attractive target for adversaries seeking to impose economic costs on a competitor. A disruption affecting a single hyperscale provider's AI inference infrastructure could cascade into thousands of enterprise applications with no obvious mitigation path for organizations that have not maintained fallback capability.

A third scenario is competitive platform divergence, in which the bifurcation of global AI infrastructure into U.S.-aligned and Chinese-aligned ecosystems forces enterprises with multinational operations to make explicit platform choices that were previously implicit. Organizations operating across the U.S. and China already navigate complex data localization and technology control requirements; as Chinese AI platforms become more capable and as regulatory environments in both countries increasingly govern which AI systems can be used for sensitive functions, enterprises may face pressure to maintain separate AI infrastructure for different operational theaters—with all the cost, complexity, and data governance implications that entails.

A fourth scenario involves provider-as-competitor dynamics escalating to the point where the enterprise's own AI vendor acquires or enters its market segment. The foundation model providers' expansion into enterprise services, managed agents, and vertical application development is already visible; as their revenue scale grows and their understanding of enterprise workflows deepens through API usage data, the competitive risk increases. This is a risk distinct from geopolitical exposure but structurally related: the enterprise's dependency on the provider makes it difficult to exit the relationship even when the provider's competitive behavior becomes adverse.

The TechNode Global analysis of geopolitical AI risk as a business continuity issue identified the underlying dynamic clearly: "When tensions rise, regulations harden, data movement gets restricted, and access to critical infrastructure, cloud regions, or strategic suppliers can change overnight. Your resilience footprint will be defined by policy shifts, sanctions, and cross-border disruption, not just malware and general vulnerabilities" [26]. This observation applies with particular force to AI, where the services are centralized, the dependencies are deep, and the policy environment is changing faster than most enterprise risk frameworks are designed to track.

A Framework for Enterprise AI Resilience

The risk environment created by frontier AI's geopolitical trajectory does not have simple solutions. Enterprises cannot readily replicate the capabilities of frontier AI providers internally, and the economic rationale for continuing to use frontier AI services remains compelling. The goal of an enterprise AI resilience framework is therefore not to eliminate dependency but to ensure that dependency is understood, bounded, and managed in ways that preserve operational continuity across a range of geopolitical scenarios.

The foundation of such a framework is comprehensive AI dependency mapping. Most enterprises do not have a complete picture of which AI services are embedded in their operational systems, which providers supply those services, where the underlying compute infrastructure is physically located, and what data flows across jurisdictional boundaries to support each AI function. Creating this inventory is the prerequisite for any risk assessment. The mapping should extend to third- and fourth-party dependencies: the AI tools embedded in SaaS applications, the AI components in development toolchains, and the AI inference used by critical suppliers all represent exposures that may not be visible in first-party vendor contracts. Enterprises should also document the operational criticality of each AI dependency—the degree to which a specific function would be disrupted by a loss of AI service access, and over what time horizon a fallback process could be implemented.

With a dependency map in place, enterprises can conduct geopolitical exposure analysis—assessing which dependencies are subject to the highest policy risk based on the provider's jurisdiction, the enterprise's operational jurisdictions, and the current regulatory trajectory in each. This analysis should be revisited at least annually given the pace of policy change observed in 2025–2026. A provider headquartered in the United States serving an enterprise with significant operations in markets that are Tier 2 under potential future export control regimes represents a different risk profile than a purely domestic deployment. Enterprises with operations in jurisdictions adjacent to active geopolitical tensions face elevated risk from both directions: from controls imposed by the AI provider's home country and from controls imposed by the host country.

Multi-provider strategy and abstraction architecture are the primary technical mitigations available at the enterprise level. The core principle is to insert an abstraction layer between application code and provider-specific APIs, so that the application's logic is written against a unified interface that can route requests to different underlying providers without application-level modification. Several commercial and open-source LLM routing frameworks have emerged specifically to address this architectural need. The operational cost of maintaining multi-provider capability—including prompt engineering investments validated across providers, model behavior testing for each provider, and the overhead of maintaining multiple provider relationships—is real, but it is the cost of maintaining meaningful provider flexibility. Enterprises that maintain the ability to route at least some workloads to a secondary provider are positioned to negotiate more effectively on pricing, to respond more rapidly to service disruptions, and to manage regulatory changes that affect access to specific providers.

Data and model portability are a related architectural consideration. Enterprises that invest heavily in fine-tuning frontier models on proprietary data should evaluate whether the resulting model weights can be extracted and hosted independently, or whether they are bound to the provider's infrastructure by the terms of the service agreement and the technical structure of the fine-tuning process. Open-weight models—those whose trained parameters are publicly available—offer a portability path that proprietary API services do not. The trade-off is that open-weight frontier models have historically trailed proprietary frontier models in capability, though the gap has narrowed considerably with models like DeepSeek R1 and its successors. Enterprises whose AI workloads can be adequately served by open-weight models gain substantially greater operational flexibility and reduced geopolitical exposure.

At the strategic level, enterprises in critical sectors—financial services, healthcare, defense contracting, energy, and critical infrastructure—should evaluate whether sovereign AI initiatives in their operating jurisdictions offer viable alternatives to U.S.-controlled frontier AI services for specific sensitive workloads. Several national AI programs are designing their compute infrastructure specifically for enterprise use in regulated sectors, with data residency guarantees and regulatory frameworks appropriate for sensitive data processing. While sovereign AI infrastructure generally lags frontier commercial providers in capability, it may be sufficient for defined workloads and offers a meaningful hedge against the geopolitical scenarios described above.

Business continuity planning for AI service disruption is a component of enterprise resilience that most organizations have not yet addressed explicitly. The same discipline applied to other critical infrastructure dependencies—identifying recovery time objectives, documenting fallback procedures, testing continuity processes—should be applied to AI services that have become operationally critical. For functions where AI service disruption would have immediate operational impact, human-staffed fallback processes should be documented and periodically exercised. Business continuity planners should work with AI operations teams to characterize disruption scenarios by provider, by workload type, and by geographic scope.

Conclusions and Recommendations

The geopolitical positioning of frontier AI as a strategic resource is not a temporary condition that will resolve as the technology matures. If anything, it intensifies as AI capability becomes more consequential for economic productivity, military effectiveness, and the functioning of government. The patterns established in 2025–2026—export controls, sovereign investment programs, bifurcating AI ecosystems, and legislative action to embed controls in hardware—reflect durable strategic interests that will shape the AI landscape for at least the remainder of this decade.

Enterprises that treat AI vendor relationships as conventional software procurement decisions are misaligned with the risk environment they actually face. AI provider access now carries a category of geopolitical risk analogous to that which enterprises apply to banking relationships in politically unstable jurisdictions or energy supply arrangements in commodity-volatile markets. The appropriate response is not paralysis or withdrawal from frontier AI, which would impose competitive costs too large to accept, but a systematic recalibration of how AI dependencies are identified, classified, and managed.

Immediate actions for enterprise security organizations include: conducting a comprehensive AI dependency inventory extending to third- and fourth-party exposures; classifying each dependency by provider jurisdiction, operational criticality, and estimated switching cost; reviewing AI vendor contracts for portability terms, data jurisdiction clauses, and service continuity provisions; and briefing senior leadership and risk committees on the geopolitical dimensions of AI vendor exposure.

Near-term priorities over the following twelve months include: implementing abstraction architecture for critical AI workloads to reduce hard coupling to specific providers; establishing a secondary provider relationship for the highest-criticality AI functions; evaluating open-weight model alternatives for workloads where portability is more important than peak capability; updating business continuity plans to address AI service disruption scenarios; and incorporating AI dependency risk into the enterprise's third-party risk management program with appropriate monitoring and review cadence.

Strategic considerations over a two-to-five-year horizon include: monitoring sovereign AI program developments in key operating jurisdictions for viable infrastructure alternatives; evaluating whether critical sector regulatory requirements in operating jurisdictions will mandate domestic or sovereign AI infrastructure for specific functions; engaging with industry bodies and regulatory processes shaping AI governance to ensure enterprise perspectives inform policy development; and building internal AI literacy in security, legal, and risk functions sufficient to assess and respond to regulatory changes with the speed those changes may require.

The organizations best positioned to navigate this environment are those that understand their AI dependencies with the same clarity they apply to other critical infrastructure, that have invested in genuine technical optionality rather than comfortable single-vendor relationships, and that treat AI governance as a board-level risk management function rather than a technical matter to be resolved in procurement negotiations. The transition from viewing AI as a technology acquisition to viewing it as a geopolitically contingent strategic asset is uncomfortable but necessary.

CSA Resource Alignment

Several CSA frameworks provide directly applicable guidance for enterprises building the risk management capabilities described in this whitepaper.

The **AI Controls Matrix (AICM)** is CSA's primary framework for AI security governance, covering supply chain security, third-party AI risk, and shared responsibility across model providers, cloud service providers, orchestrated service providers, and AI customers. Enterprises conducting AI dependency mapping and vendor risk assessments should reference AICM's supply chain security domain for control requirements applicable to frontier AI provider relationships. The AICM's shared security responsibility model provides a structured approach for allocating security obligations between enterprises and their AI providers—an allocation that has direct relevance when assessing exposure to geopolitically driven service changes.

The **STAR for AI** program enables AI service providers to demonstrate security assurance through Level 1 self-assessment (AI-CAIQ) submissions to the CSA STAR Registry. Enterprises evaluating frontier AI providers as part of vendor risk programs should check provider STAR for AI certifications as an indicator of security transparency; absence of such certification is relevant context for provider risk classification, particularly for high-criticality workloads.

The **AI Organizational Responsibilities: Governance, Risk Management, Compliance** publication provides guidance on AI GRC frameworks applicable to the strategic governance requirements identified in this paper, including board-level AI risk reporting, shadow AI governance, and AI incident response planning. Its frameworks for AI risk management strategy are directly applicable to the multi-year resilience planning recommended above.

The **MAESTRO** framework for agentic AI threat modeling applies to enterprises deploying AI agents that depend on frontier model APIs. MAESTRO's threat modeling approach can be adapted to assess the cascading impacts of AI provider service disruption on agentic workloads where operational continuity depends on uninterrupted API access.

CSA's **Zero Trust** guidance addresses the architectural principles applicable to multi-provider AI environments, where trust cannot be assumed at the network or provider level and must be established through policy-enforced access controls and continuous verification. Zero trust principles applied to AI infrastructure support the abstraction and portability strategies recommended in this paper.

Additional CSA resources relevant to this domain include the **Cloud Controls Matrix (CCM)** supply chain management controls, which provide a complementary control catalog for assessing AI provider relationships within existing cloud governance programs.

References

[1] World Economic Forum. "Global Cybersecurity Outlook 2026." World Economic Forum, January 2026.

[2] U.S. Department of Commerce. "Framework for Artificial Intelligence Diffusion." Federal Register, January 15, 2025.

[3] Covington & Burling. "U.S. Department of Commerce Establishes Export Control Framework Limiting the Diffusion of Advanced Artificial Intelligence." Covington & Burling, January 2025.

[4] Kirkland & Ellis. "BIS Rescission of the Biden Administration's AI Diffusion Framework." Kirkland & Ellis, May 2025.

[5] Morgan Lewis. "BIS Revises Export Review Policy for Advanced AI Chips Destined for China and Macau." Morgan Lewis, January 2026.

[6] The Cyber Express. "Congress Wants A Tracker On Every Advanced AI Chip US Exports." The Cyber Express, 2026.

[7] Atlantic Council. "How the Chip Security Act Could Usher in an Era of 'Trusted Trade' with US Partners." Atlantic Council, 2026.

[8] RAND Corporation. "Understanding the Artificial Intelligence Diffusion Framework: Can Export Controls Create a U.S.-Led Global Artificial Intelligence Ecosystem?" RAND Corporation, 2025.

[9] Titan Investors. "Sovereign Wealth Funds Commit $120 Billion to AI Infrastructure Buildout." Titan Investors, 2026.

[10] Compute Forecast. "Sovereign AI Infrastructure Is Now a Nation-State Race." Compute Forecast, 2025–2026.

[11] Introl. "South Korea's $735B Sovereign AI Initiative." Introl, 2026.

[12] European Commission. "Strengthening Europe's Tech Sovereignty." European Commission, June 3, 2026.

[13] European Commission. "Commission Proposes Tech Sovereignty Package to Strengthen Europe's Digital Autonomy and Resilience." European Commission Press Corner, 2026.

[14] Tech Policy Press. "Europe's AI Sovereignty Problem Runs Far Deeper Than Frontier Access." Tech Policy Press, 2026.

[15] International Institute for Strategic Studies. "Sovereign AI: Pathways to Strategic Autonomy." IISS, August 2025.

[16] McKinsey & Company. "Sovereign AI Ecosystems for Strategic Resilience and Economic Impact." McKinsey Global Institute, 2025–2026.

[17] European Union Institute for Security Studies. "Challenging US Dominance: China's DeepSeek Model and the Pluralisation of AI Development." EUISS, 2025.

[18] Council on Foreign Relations. "DeepSeek V4 Signals a New Phase in the U.S.-China AI Rivalry." CFR, 2026.

[19] OgunSecurity. "DeepSeek V4 Release: China's Sovereign AI Stack and the Strategic Fracturing of US Technology Dominance." OgunSecurity, 2026.

[20] U.S.-China Economic and Security Review Commission. "Two Loops: How China's Open AI Strategy Reinforces Its Industrial Dominance." USCC, March 2026.

[21] CSIS. "DeepSeek, Huawei, Export Controls, and the Future of the U.S.-China AI Race." Center for Strategic and International Studies, 2025.

[22] Cloud Pro Inc. "AI Vendor Lock-In: How to Manage Risk with Anthropic, OpenAI, and Google." Cloud Pro, April 2026.

[23] Zapier. "Zapier Survey Finds Nearly 3 in 4 Enterprises Would Face Disruption If They Lost Their Primary AI Vendor." BusinessWire, April 2, 2026.

[24] CIO. "OpenAI, Anthropic Expand Services Push, Signaling New Phase in Enterprise AI Race." CIO, 2026.

[25] HSToday. "Critical Infrastructure and the Rising Concentration Risk in Digital Ecosystems." Homeland Security Today, 2025.

[26] TechNode Global. "Why Geopolitical Conflict Is Making AI Risk a Business Continuity Issue." TechNode Global, March 2026.

[27] Brookings Institution. "Competing AI Strategies for the US and China." Brookings Institution, 2025–2026.

[28] Mayer Brown. "Administration Policies on Advanced AI Chips Codified, with Reverberations Across AI Ecosystem." Mayer Brown, January 2026.

[29] NPR. "Trump Administration Grants NVIDIA License to Sell H20 Chips to China With Revenue-Sharing Condition." NPR, August 11, 2025.