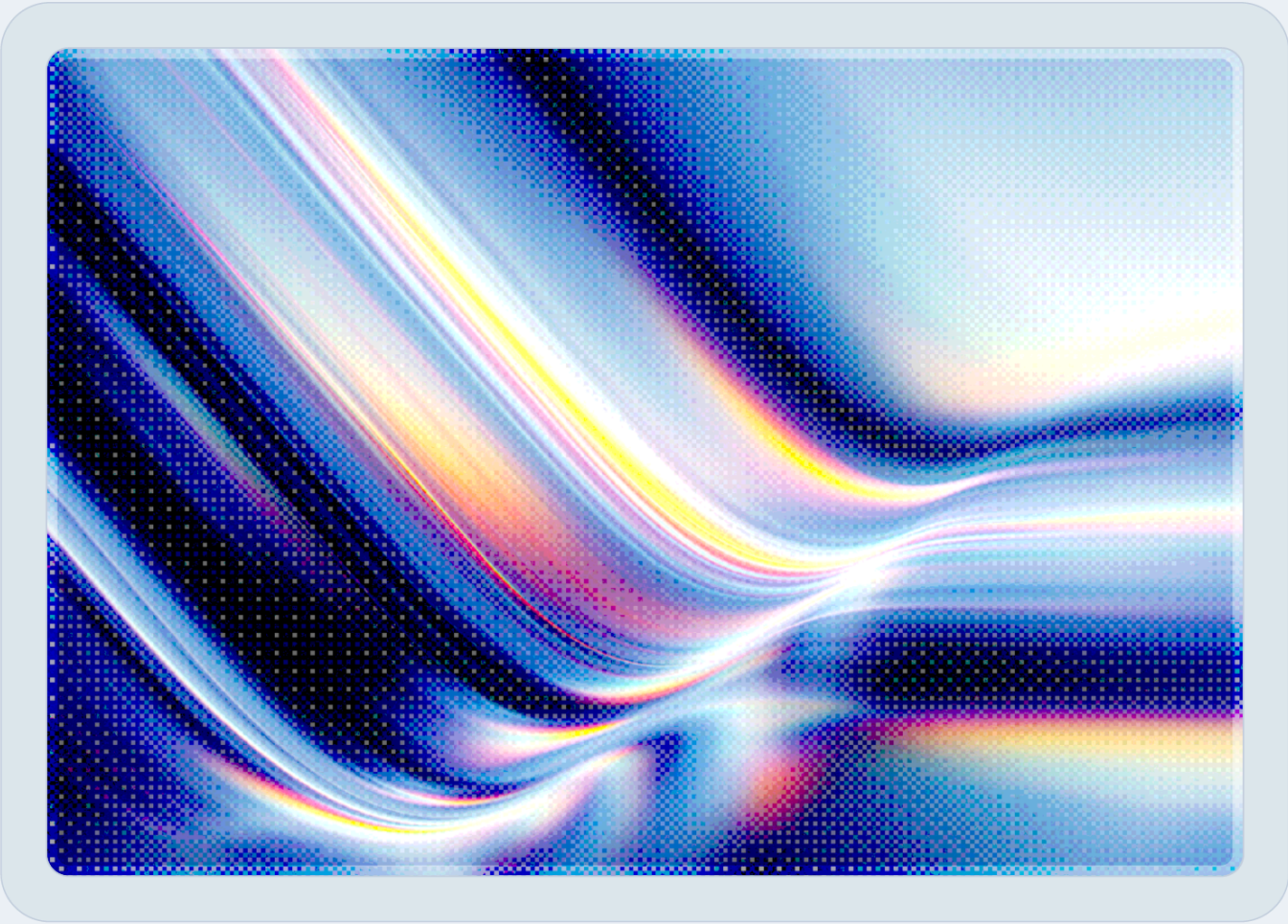
CSAI Foundation | Cloud Security Alliance

Sovereign Dependency Risk: One AI Vendor, National Infrastructure

2026-09-25

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- Anthropic's Claude became, by public accounts, the first large language model deployed inside U.S. classified networks, running under a Department of War prototype agreement worth up to $200 million and reportedly supporting intelligence analysis, operational planning, and cyber operations at classification levels up to Top Secret [1][2]. That depth of integration turned a single commercial vendor into a functional piece of national security infrastructure well before policymakers had settled how to govern that dependency.
- After Anthropic resisted a Pentagon contract term permitting "any lawful use" of Claude, citing risks tied to autonomous weapons and domestic surveillance, Defense Secretary Pete Hegseth and President Trump moved in late February 2026 to brand the company a "supply-chain risk to national security" and direct all federal agencies to stop using its products [3][4].
- A federal district judge found in March 2026, and again in a final ruling on August 27, 2026, that the designation and the government-wide ban were unlawful, calling the government's actions "classic illegal First Amendment retaliation" and a due-process violation, while a parallel case remained active in Washington and the government retained the option to appeal [5][6].
- A separate June 2026 export-control action by the Bureau of Industry and Security led Anthropic to voluntarily suspend global access to two newly released models, Fable 5 and Mythos 5, for nineteen days rather than risk unverifiable partial compliance, illustrating that a single vendor's frontier models can become unavailable to every customer worldwide as a side effect of a regulatory mechanism entirely unrelated to any contract dispute [7][8].
- Together, these episodes show that sovereign dependency risk is no longer speculative: it has already produced a government-directed vendor ban, a global model shutdown, and litigation over whether an AI company that declines a government's terms can be branded a national security threat, all within a single AI vendor's product line inside of eight months.

Background

Frontier AI vendors did not become national security infrastructure gradually and by policy design; they became infrastructure quickly, through a small number of procurement decisions layered on top of each other. In July 2025, the Department of Defense's Chief Digital and Artificial Intelligence Office awarded

parallel prototype agreements worth up to $200 million each to Anthropic, OpenAI, Google, and xAI, explicitly intended to prevent single-vendor lock-in by cultivating four competing suppliers rather than one [1][9]. That multi-vendor structure looked, on paper, like a standard concentration-risk mitigation. In practice, the four vendors were not interchangeable. Anthropic alone had already made Claude available on a Palantir-hosted platform cleared to handle information up to the Secret level in late 2024, and its subsequent Claude Gov models were built to run inside classified, on-premises government cloud environments, including Amazon Web Services' offerings for Top Secret workloads [2][9]. By early 2026, Anthropic's models were, according to the company's own public statements, in use across the Department of War and other national security agencies for intelligence analysis, modeling and simulation, operational planning, and cyber operations, a depth of embedding in classified environments that, based on public reporting, OpenAI, Google, and xAI had not yet matched [1][9].

That asymmetry mattered once Anthropic and the Pentagon reached an impasse over how Claude could be used. During contract renegotiations in late 2025 and early 2026, the Department of Defense sought a term permitting "any lawful use" of Claude, and Anthropic declined, citing its own acceptable-use restrictions on autonomous weapons targeting and domestic mass surveillance conducted without human oversight or judicial process [3][10]. Anthropic's chief executive made the disagreement public in a statement on February 26, 2026. Within days, Hegseth and Trump responded, not through a negotiated contract amendment but through public directives: Hegseth designated Anthropic a "supply-chain risk to national security," a label that federal agencies have used in other contexts to support excluding a company from government contracts, and Trump ordered all federal agencies to cease using Anthropic's technology [3][4]. That response arrived against a backdrop already in motion rather than one it set off: Pentagon officials had confirmed a parallel agreement to bring xAI's Grok into classified systems on February 24, 2026, two days before Anthropic's public disclosure and before the designation itself, so the xAI arrangement was not a rapid substitution manufactured in response to the Anthropic ban but a deal already under way — one that, according to contemporaneous reporting, nonetheless signaled that willingness to accept looser use restrictions, rather than technical capability alone, had begun to shape which vendor gained access to the most sensitive government workloads [10][11].

Anthropic did not accept the designation quietly. In early March 2026 the company filed suit against the Trump administration, arguing that the supply-chain-risk label and the governmentwide ban were retaliation for its public stance on AI safety rather than a genuine security determination [4][5]. U.S. District Judge Rita Lin agreed with the essential thrust of that argument twice. On March 26, 2026, she granted a preliminary injunction blocking both the ban and the designation, writing that "punishing Anthropic for bringing public scrutiny to the government's contracting position is classic illegal First Amendment retaliation" [5]. The government appealed and continued to press its position through the spring, but on August 27, 2026, Judge Lin issued a final ruling vacating the designation altogether, finding it "arbitrary and capricious" under the Administrative Procedure Act and concluding that Anthropic had been denied the pre-deprivation process the Fifth Amendment requires [6][12]. A second,

related Anthropic suit remained active in a Washington, D.C. federal court as of that ruling, and the Pentagon retained the ability to appeal the California decision, so the underlying legal question of how far a government agency can go in branding a compliant, operating vendor a security threat had not been fully and finally resolved even after six months of litigation [6][12].

Security Analysis

This analysis treats the mechanics of the Pentagon-Anthropic dispute as more instructive than its ultimate outcome. Anthropic held to its acceptable-use restrictions against enabling autonomous lethal targeting and unchecked domestic surveillance, and the district court ultimately characterized the government's response to that position as unlawful retaliation [5][6]. That finding illustrates that a government customer unhappy with a vendor's use restrictions has policy tools well beyond ordinary contract non-renewal at its disposal, including a national-security-risk designation more commonly associated with foreign adversary technology and a public order directing every federal agency to sever ties with a single company. Those tools moved faster than the litigation needed to test their legality, meaning that for the roughly six months between Hegseth's directive and Judge Lin's final ruling, Anthropic and any agency relying on its models operated under genuine legal uncertainty about whether Claude access would remain available at all. The court record in this case indicates that an organization's own contractual guardrails, however well designed, cannot fully insulate it from a government counterparty's capacity to use procurement and export authorities against a vendor whose policies it dislikes, a structural vulnerability CSA's own research on this incident has previously identified [13].

The near-simultaneous export-control episode involving Fable 5 and Mythos 5 shows the same underlying risk from a different angle, with no contract dispute or policy disagreement involved at all. On June 12, 2026, the Bureau of Industry and Security issued a directive under the Export Control Reform Act requiring individually validated export licenses before any foreign national, including Anthropic's own foreign national employees, could access either model, invoking the "deemed export" doctrine that treats providing model access to a foreign person as itself a controlled export [7][8]. Because Anthropic could not verify user nationality in real time across the dozens of cloud platforms and integrations through which the models were delivered, it disabled both models globally rather than risk partial, unverifiable compliance. Fable 5 and Mythos 5 remained offline for nineteen days before the Commerce Department lifted the controls on June 30, with Fable 5 restored to general availability on July 1 and Mythos 5 reintroduced on a more limited basis to vetted U.S. organizations [14][15]. No standard enterprise contract, service-level agreement, or acceptable-use negotiation in place at the time appears to have prevented that outage; it originated in a regulatory authority that operates independently of, and faster than, ordinary commercial dispute resolution. The two episodes together, one arising from a

contract disagreement and one from export-control enforcement, demonstrate that a single vendor's models can become unavailable through at least two structurally distinct mechanisms that most enterprise and government risk registers do not yet treat as a single category of exposure.

This concentration risk is compounded by what a single vendor's classified footprint means for the broader national security enterprise. Where a commercial customer facing a vendor dispute or export restriction can typically switch to a substitute product within weeks or months, an agency running intelligence analysis or operational planning workflows on a model qualified for Top Secret environments faces qualification, accreditation, and retraining costs that make rapid substitution far harder. The Pentagon's parallel classified-systems arrangement with xAI, already reported on February 24, 2026 and therefore in motion before rather than as a rapid response to the Anthropic ban, nonetheless illustrates the risk running in the opposite direction: concentrating classified national security workloads on a second vendor, selected in part, according to contemporaneous reporting, because it agreed to fewer use restrictions [10][11], does not resolve sovereign dependency risk so much as relocate it, while potentially trading away the safety guardrails the original dispute was about. Meanwhile, the White House's June 5, 2026 National Security Presidential Memorandum on AI in the national security enterprise, NSPM-11, explicitly directs agencies to "maintain deep, proactive partnerships with industry" to accelerate adoption and calls for ensuring that deployed AI systems "cannot be disabled or altered without federal government approval," a goal the Fable 5 and Mythos 5 shutdown shows the government does not yet reliably control even for models it did not order disabled on that occasion [16][17].

Mechanism	Trigger	Who Controls the Off-Switch	Duration Observed
Government-directed vendor ban	Contract/policy disagreement (Pentagon "any lawful use" clause)	Executive branch and agency officials	~6 months in litigation; designation ultimately vacated [5][6]
Export-control directive	National-security classification of model access ("deemed export")	Commerce Department / Bureau of Industry and Security	19 days (June 12–30, 2026) [7][14]
Vendor-initiated substitution	Customer declines vendor's use restrictions	Vendor and competing suppliers	Parallel arrangement already reported before the ban [10][11]

Recommendations

Immediate Actions

Agencies and enterprises operating AI workloads that touch national security functions, critical infrastructure, or other mission-critical processes should inventory which vendors' models are embedded deeply enough, through classification-level accreditation, custom fine-tuning, or workflow integration, that substitution would take months rather than weeks, and should treat that subset of dependencies as requiring board- or agency-leadership-level visibility rather than routine vendor management. Legal and contracts teams should review existing AI vendor agreements for force majeure and regulatory-disruption clauses, since standard commercial force majeure language may not have anticipated export-control-driven, all-customer model shutdowns of the kind seen in the Fable 5 and Mythos 5 episode.

Short-Term Mitigations

Organizations with classified or otherwise highly sensitive AI workloads should pursue qualification and accreditation of at least one alternative model or vendor in parallel with their primary deployment, even where that primary vendor is currently performing well, so that a government-directed ban, an export-control action, or a unilateral vendor decision does not leave mission-critical functions without a tested fallback. Where models are deployed for functions the government considers essential, agencies should press for, and vendors should offer, contractual or architectural commitments around advance notice and phased degradation in the event of a government-directed access restriction, rather than the all-or-nothing global shutdown that characterized the June 2026 export-control response.

Strategic Considerations

Both government and enterprise customers should recognize that vendor "diversification" achieved by contracting with multiple frontier AI providers, as the Department of Defense did in 2025, does not by itself eliminate concentration risk if only one of those vendors is actually accredited and embedded at the classification level or operational depth that matters for a given mission. Genuine resilience requires validating that alternative vendors are qualified to the same operational standard as the primary vendor, not merely under contract. Policymakers should also weigh the second-order effects of using national-security-risk designations or export-control authorities as leverage in ordinary commercial and policy disputes with AI vendors: doing so establishes a precedent, now subject to ongoing litigation, that could

equally be turned against any vendor whose safety commitments conflict with a future administration's priorities, a dynamic that ultimately concentrates power over widely used AI infrastructure in whichever administration currently holds the relevant authorities rather than reducing dependency risk itself.

CSA Resource Alignment

This incident sits at the center of CSA's own research note Pentagon vs. Anthropic: Autonomous Weapons AI Guardrails and the Governance Crisis for Enterprise AI Vendors, published as the dispute first became public, which concluded that contractual acceptable-use policies, while necessary, are insufficient as primary controls once a government counterparty has the leverage and motivation to override them, and that the supply-chain-risk designation transformed an ordinary vendor governance disagreement into a coercive mechanism with implications reaching well beyond defense contracting. The final court ruling analyzed in this note confirms that report's central concern: the designation was found unlawful, but only after roughly six months during which the coercive mechanism itself was live and operative.

CSA's AI Developer Ecosystem Concentration: Critical Infrastructure's Hidden Risk provides the structural framing for why this single-vendor episode matters beyond Anthropic and the Pentagon. That research documents how consolidation among a small number of foundation model providers has crossed a threshold at which vendor concentration should be treated as a critical infrastructure problem rather than an ordinary vendor management question, and it specifically warns that a single exploited or disrupted vendor can cascade risk across every institution that depends on it simultaneously, an instance of that same dynamic, illustrated when Fable 5 and Mythos 5 went offline for every customer worldwide over one regulatory action.

CSA's Sovereign AI Risk: When Your AI Vendor Gets Export-Controlled offers a complementary governance framework for enterprises and agencies seeking to operationalize the lessons of both episodes, including guidance on architecting nationality-aware and jurisdiction-aware access controls of the kind that might have mitigated, though likely not eliminated, the impact of the June 2026 export-control shutdown. Finally, the identity, supply-chain, and business-continuity domains of the AI Controls Matrix (AICM v1.1) supply the auditable control language organizations need to translate "know which vendors you cannot quickly replace" from a strategic observation into a documented, assessable governance practice.

References

[1] Anthropic. "Anthropic and the Department of Defense to advance responsible AI in defense operations." Anthropic, July 2025.

[2] Anthropic. "Claude in Amazon Bedrock: Approved for use in FedRAMP High and DoD IL4/5 workloads." Anthropic, 2025.

[3] NPR. "Anthropic sues the Trump administration over 'supply chain risk' label." NPR, March 9, 2026.

[4] Al Jazeera. "Anthropic sues Trump administration to undo US 'supply chain risk' tag." Al Jazeera, March 9, 2026.

[5] CNBC. "Anthropic wins preliminary injunction in DOD fight as judge cites 'First Amendment retaliation'." CNBC, March 26, 2026.

[6] CNN Business. "Judge rules the Pentagon's supply chain risk label for Anthropic unlawful." CNN, August 27, 2026.

[7] Digital Applied. "Fable 5 & Mythos hit US export controls: what it means." Digital Applied, June 20, 2026.

[8] Forbes. "Anthropic disabled Fable 5 and Mythos 5 after a U.S. export-control order. Here's what happened." Forbes, June 16, 2026.

[9] IBTimes UK. "OpenAI, Google, Anthropic and xAI's up-to-$200M Pentagon deals help shape military AI strategy." IBTimes UK, 2025.

[10] Teslarati. "xAI's Grok approved for Pentagon classified systems: report." Teslarati, February 2026.

[11] Anews. "Pentagon, Musk's xAI reach agreement to use Grok in classified systems." Anews, February 24, 2026.

[12] CNBC. "Judge blocks Pentagon blacklist of Anthropic as supply chain risk." CNBC, August 28, 2026.

[13] Cloud Security Alliance. "Pentagon vs. Anthropic: Autonomous Weapons AI Guardrails and the Governance Crisis for Enterprise AI Vendors." Cloud Security Alliance, March 9, 2026.

[14] CNBC. "Anthropic says Trump admin has lifted export controls on Claude Fable 5 and Mythos 5." CNBC, June 30, 2026.

[15] MarketScale. "Fable 5 and Mythos 5 are back. What the 19-day shutdown taught every enterprise about AI as infrastructure." MarketScale, July 2026.

[16] The White House. "National Security Presidential Memorandum/NSPM-11." The White House, June 5, 2026.

[17] Everfox. "How NSPM-11 is shaping AI for national security." Everfox, 2026.