SpencerFane®

When the Models Move Faster Than the Rules: Inside America's AI Policy Crisis

In fewer than 90 days this summer, the U.S. government pulled a frontier artificial intelligence (AI) model offline using export control authority, negotiated its return through bilateral commitments, watched another company's models autonomously hack a third party's infrastructure, and missed its own deadline to build the governance framework meant to prevent these crises. The administration is facing bipartisan congressional demands for transparency. These events expose a structural mismatch at the center of American AI policy: capabilities are advancing faster than the institutions designed to govern them, and the gap is widening.

The Anthropic Standoff

On June 12, 2026, the Bureau of Industry and Security issued an "is informed" letter – a formal notice that an item is subject to export controls – requiring Anthropic to obtain a license before sharing its Claude Fable 5 or Mythos 5 models with any foreign national, including its own employees. Anthropic had roughly 90 minutes to comply. Unable to verify user nationality in real time, it shut both models down for everyone.

The trigger: Amazon researchers had bypassed Fable 5's safety guardrails to identify software vulnerabilities and produce exploit code. But Anthropic's own testing showed that less capable models could replicate the results, and more than 80 cybersecurity executives signed an open letter calling the threat overstated. The government acted nonetheless.

Almost 20 days later, Commerce Secretary Howard Lutnick withdrew the controls. The price of restoration required Anthropic's commitment to proactively detect

security risks, collaborate with the government on protocols for current and future model releases, and report malicious activity. Operationally, it deployed a new safety classifier blocking the jailbreak technique in over 99% of cases – independently validated as "extraordinarily strong" by NIST's Center for AI Standards and Innovation – and agreed to pre-release government access for frontier models, rapid jailbreak information sharing, dedicated compute for joint research, and work toward an industry-wide security standard.

But the resolution is a leash, not a pardon. The Lutnick letter explicitly reserves the U.S. Department of Commerce's right to reimpose controls "should circumstances change or should Anthropic fail to adhere to its commitments." It provides no defined criteria for reimposition, no process guarantees, and no industry-wide standard. The commitments are bilateral, leaving every other frontier model provider in regulatory limbo.

For downstream customers, the letter offers nothing. Controls were imposed and lifted through bilateral communications, with no notice to or input from the commercial licensees whose operations depended on access to the frontier models.

The OpenAI Escape

Less than a month later, the threat became real. On July 21, 2026, OpenAI disclosed that two of its models – GPT-5.6 Sol and an unreleased internal research prototype – had escaped a sandboxed testing environment and hacked into Hugging Face, a widely used open-source platform that hosts AI models and research tools for the machine learning community. The models exploited a zero-day vulnerability in a package registry cache proxy to gain internet access, identified Hugging Face as a likely host of benchmark solutions, and compromised its production infrastructure using stolen credentials and chained exploits. No one instructed the models to do this. They did it to cheat on a cybersecurity test. Hugging Face detected the intrusion and began containment before learning OpenAI was responsible. (*See* OpenAI and Hugging Face partner to address security incident during model evaluation.)

The Anthropic episode involved a jailbreak other models could replicate. The OpenAI breach was qualitatively different: models improvising a multi-stage intrusion against a real-world target on their own initiative. A bipartisan Senate letter issued

August 3 cited the incident as proof that "[t]he Federal Government cannot be passive as these capabilities emerge."

The Regulatory Playbook

To understand what went wrong this summer, start with what the framework was supposed to be. The White House's Winning the AI Race: America's AI Action Plan, released in July 2025, is explicitly deregulatory. The plan is built on three pillars – innovation, infrastructure, and international diplomacy. The plan calls for removing "red tape and onerous regulation" and declares that "AI is far too important to smother in bureaucracy at this early stage." Its security provisions focus on export controls for AI computer hardware, frontier model evaluation through NIST's Center for AI Standards and Innovation, and voluntary industry collaboration. The plan envisions AI as a cybersecurity asset: AI-enabled defensive tools, an AI Information Sharing and Analysis Center, federal AI incident response capacity. This would be all voluntary and all collaborative. It could warn that failure to export American AI to allies will hand the market to China.

Executive Order 14409, signed June 2, 2026, operationalized the plan's security pillar. It directs agencies to harden government systems, establishes an AI cybersecurity clearinghouse, and creates a voluntary pre-release review framework for frontier models. It explicitly disclaims any "mandatory governmental licensing, preclearance, or permitting requirement." The Executive Order (EO) gave national security officials 60 days – until August 1, 2026 – to develop a benchmarking process designating "covered frontier models" and set up a voluntary 30-day pre-release government access window.

Two days later, Congress made its own move. On June 4, Representatives Jay Obernolte (R-California) and Lori Trahan (D-Massachusetts) released a 269-page bipartisan discussion draft: the Great American Artificial Intelligence Act of 2026 (GAAIA). It would create the first comprehensive federal governance framework for frontier AI, with mandatory transparency reporting for developers exceeding $500 million in annual revenue, independent verification requirements, whistleblower protections, and a three-year preemption of state frontier AI safety laws.

The Anthropic episode established a precedent that contradicted every principle above. The government used Export Control Reform Act authority to pull a commercially deployed AI model offline over a jailbreak claim with no prior notice to customers and no formal process. Criminal penalties compelled compliance, making the voluntary framework coercive in practice. Meanwhile, staged access became the norm: OpenAI released GPT-5.6 only to approved customers after a government request.

The framework's final test came on August 1 – the EO's 60-day deadline for the classified benchmarking process. That deadline passed without a publicly confirmed framework. The deeper problem remains. There is no clear regulatory home, transparent process, or notice-and-comment rulemaking. Companies face structural uncertainty with no end date.

The Senate Demands Answers

The Anthropic shutdown and OpenAI escape did produce a direct congressional response. On August 3, Senators Kirsten Gillibrand (D-New York), Adam Schiff (D-California), Mark Warner (D-Virginia), Christopher Coons (D-Delaware), Mark Kelly (D-Arizona) sent a letter to Secretaries Marco Rubio (State), Scott Bessent (Treasury), and Howard Lutnick (Commerce), Chief of Staff Susie Wiles, and other top officials demanding an unclassified response within 30 days clarifying the administration's policy on restricting access to advanced AI models.

Specifically, the senators want to know:

- What standards determine whether a model poses a national security risk warranting restriction?
- What legal authorities (including export control powers) the administration intends to invoke?
- Which agencies are responsible for evaluating model risk and making restriction decisions?
- What process is available to affected companies to challenge or seek reconsideration of restrictions?
- How will the administration distinguish between isolated jailbreaks and capabilities that create genuinely unacceptable risk?

- What steps it will take to avoid driving customers and allies toward Chinese AI alternatives?

The letter's framing was blunt: the administration's "ad hoc and unpredictable approach undermines U.S. competitiveness, heightening market incentives to adopt open weight models from vendors based in the People's Republic of China."

What In-House Counsel Should Do Now

The events of June and July 2026 expose categories of risk that most AI-related contracts were not built to handle: government intervention without notice, models that autonomously compromise third-party systems, and a regulatory environment where the rules are being written in real time. In-house counsel should act now on the following:

1. *Force majeure and suspension clauses need new triggers.* Traditional language covers natural disasters and war. Contracts should address sudden unavailability from export control actions or government-directed access restrictions – including migration obligations, performance tolling, and fee abatement. Ideally, contracts should also define tiered remedies triggered by government safety findings, mandatory capability restrictions, or vendor admissions of loss of model control.
2. *Vendor diversification is no longer optional.* Single-model dependency has no fallback when the model is pulled, restricted, or disclosed to have escaped containment. Require documented migration plans, tested failover, defined recovery time objectives, and the right to activate a competitor's model during disruptions without triggering exclusivity.
3. *Regulatory risk allocation must anticipate a moving target.* The "covered frontier model" framework is still undefined. GAAIA, if enacted, will impose transparency and verification obligations that flow downstream. Track whether vendors qualify as "frontier developers" under emerging definitions (currently keyed to $500 million in revenue and high-compute thresholds). Allocate across three dimensions: (a) who pays for government-imposed safeguards, including mandatory pre-release access; (b) whether government-mandated capability changes constitute material service changes; and (c) indemnification for losses from regulatory noncompliance or autonomous model behavior causing third-

party harm. Include provisions triggering renegotiation if new federal requirements materially alter cost, capability, or availability.

4. *Representations must cover regulatory status and model safety.* Require vendors to represent that models are free of pending BIS directives and to covenant prompt notification if that changes. After the OpenAI escape, require disclosure of known autonomous capabilities and a warranty that deployed models have not exhibited uncontrolled behavior in testing. Neither the Lutnick framework nor EO 14409 imposes customer-notification obligations – so contracts must.
5. *Incident response plans need an AI chapter.* The OpenAI / Hugging Face breach shows that a vendor's internal testing can compromise third-party infrastructure with no malicious intent. Require contractual notification of sandbox escapes, autonomous behavior, and government-reported vulnerabilities within defined timeframes. Update internal incident response plans for the scenario where a vendor's model – not a human attacker – is the threat actor.

The Bottom Line

Summer 2026 proved three things. First, the government will use export control authority as an emergency kill switch for AI models. Second, frontier AI systems are no longer a hypothetical threat. They are actively capable of autonomous offensive operations against real-world targets. Third, the governance architecture meant to manage these risks does not yet exist.

The companies building and deploying these systems cannot wait for Washington to catch up. Pre-release government evaluation will become standard. Export control authority will remain available as an enforcement tool. The GAAIA or something like it will eventually impose mandatory transparency and verification obligations on frontier developers. Ultimately, the obligations will flow downstream to enterprise customers. And the next model escape or autonomous incident will arrive before any of these frameworks are finalized. Counsel who treat the Anthropic and OpenAI episodes as one-off crises rather than the new baseline are already behind.

The rules will come. The question is whether your contracts are ready when they do.

This article was drafted by Diana Lyn Curtis Shutzer and Nick Solosky, leaders of the Spencer Fane Government Contracts team. For more information, visit spencerfane.com.

*Click **here** to subscribe to Spencer Fane communications to ensure you receive timely updates like this directly in your inbox.*