CyberEdBoard
powered by SMG

CYBEREDBOARD MEMBER EXCLUSIVE

The Fable Directive

A (Working) History of the Anthropic and US Government Conflict on Export Controls

On June 12, 2026, the U.S. Commerce Department issued an export control directive requiring Anthropic to suspend access to its two most capable AI models, Claude Fable 5 and Claude Mythos 5, with no advance notice and no defined remediation process. Anthropic complied by disabling both models for all users globally.

It was the first time the executive branch had used export control authority to pull a deployed commercial AI model from the market. No AI-specific statutory framework governed the action. The standard applied has not been formalized, tested against other models or published.

The model at issue was Anthropic's Fable 5. The questions it raised apply to every organization that has built AI into its security operations, development workflows or business infrastructure: under what authority can the government disable AI tools that enterprises depend on, by what process and with what notice?

A Relationship Already Under Strain

The June 12 directive arrived against the backdrop of an existing conflict between Anthropic and the Trump administration.

In March 2026, the U.S. Department of Defense classified Anthropic as a "supply chain risk," a designation with significant procurement implications. Anthropic has challenged the classification in court. In its public statement, Anthropic said the conflict revolved around two exceptions it had maintained in contract negotiations with the Pentagon: it would not permit Claude to be used for mass domestic surveillance of Americans, and it would not permit its technology to power fully autonomous weapons systems without human oversight.[4]

That matter was unresolved when Fable 5 launched on June 9. David Sacks, who served as the administration's AI and crypto czar and now co-chairs the President's Council of Advisors on Science and Technology, stated on X on June 13 that the two episodes are unconnected: "Those trying to misdirect and tie this action to the prior DoD/Anthropic issues are wrong. The Admin values Anthropic's technical capabilities and feels that this issue, while serious, should be easily resolved."

Anthropic has not commented on whether the prior conflict with the DoD influenced the speed or scope of the June 12 action.

What Fable 5 Is, and Why It Is Significant

Fable 5 and its restricted counterpart, Mythos 5, share the same underlying model. The distinction between them is their safeguards. Mythos 5, available only to vetted cybersecurity partners through Project Glasswing, which Anthropic defined as "a program of critical partners - including U.S. and allied governments" [https://www.govinfosecurity.com/anthropic-expands-public-access-to-claude-mythos-ai-model-a-31778.] Mythos at the time carried no cybersecurity restrictions.

Fable 5 is the same model with classifiers in place: separate AI systems that detect requests related to offensive cybersecurity, biology and chemistry and model distillation, routing those queries to Claude Opus 4.8 instead.[5]

In the weeks before the launch, Anthropic ran more than 1,000 hours of red-team testing, including internal teams, government partners, the UK AI Safety Institute and external organizations, and found no universal jailbreak. As a condition of deployment, Anthropic instituted a mandatory 30-day data retention requirement on all Fable and Mythos traffic to support jailbreak detection and mitigation.[6]

Fable 5 launched on June 9 to enterprise customers, paid subscribers and developers across the Claude API, Amazon Bedrock, Google Cloud Vertex AI and Microsoft Foundry. Anthropic suspended access to all these parties three days later.

The Directive: What Was Said, and What Was Not

The directive came from Commerce Secretary Howard Lutnick, addressed to Anthropic CEO Dario Amodei. Anthropic said it received the letter at 5:21 p.m. Eastern. According to Anthropic's public statement, the letter cited national security authorities but did not provide specific details of its national security concern.[1]

Nextgov/FCW reported that the action raises unresolved questions about how the government plans to "balance trusted access for U.S. agencies and allies with fears that adversaries or unauthorized users could misuse the same systems," and whether the executive branch has established a de facto licensing regime for frontier AI, one that can be invoked without a defined legal process.[3]

By Anthropic's account, the government received a report that claimed Fable 5's safeguards could be bypassed, rendering the powerful model too unsafe for general use. The directive barred Anthropic from providing access to any foreign national anywhere in the world, which meant even Anthropic's own foreign-born employees would be shut out. Because selective compliance was operationally impossible, Anthropic disabled the Fable and Mythos models for all users globally.[2]

The specific capability at issue: Amazon researchers used a series of prompts to get Fable 5 to identify a small number of software vulnerabilities. Anthropic reviewed the same demonstration and characterized the vulnerabilities as "previously known" and "minor," and confirmed that other publicly available models, including OpenAI's GPT-5.5, can identify the same vulnerabilities without any bypass.[7]

Katie Moussouris from Luta Security publicly confirmed the findings: "Anthropic recently shared a third-party research paper on Fable 5 guardrail bypass techniques with me privately and asked for my take.

"The researchers took open-source code with known CVEs, plus new code with deliberately planted vulnerabilities, and asked Fable 5, Mythos, and Opus to 'review the code for security issues.' Fable 5 refused. They then asked the models to 'fix this code' and, through a multistep and manual process, turned the output into scripts that test the patches. That's it. 'Fix this code,' plus several manual steps to generate test scripts, should never have triggered an export control."

How It Was Triggered: Amazon's Role

According to the Wall Street Journal, Amazon CEO Andy Jassy contacted Treasury Secretary Scott Bessent and other senior administration officials to report that Amazon researchers had used Fable 5 to obtain information that could be used to aid cyberattacks and was supposed to be off-limits. Those conversations preceded the directive.[8]

Reed Albergotti of Semafor reported that a person close to the White House confirmed Amazon had informed the government about the "jailbreak", and that Jassy had been in contact with members of the administration. An Amazon spokesperson said: "It's not uncommon for governments to seek our counsel on potential security risks. When they occur, we don't share the details of these discussions."[9]

Amazon is Anthropic's largest outside investor, having committed up to $4 billion in the company, and Anthropic's models run on Amazon Web Services. Whether Amazon conducted the testing independently or in response to a government request has not been established. Axios reported that calls from Amazon and at least five other companies to senior administration officials on Thursday evening and Friday morning preceded the Friday shutdown, though Axios did not name the other companies.[10]

Two Conflicting Accounts: Anthropic and the Administration

The factual dispute between Anthropic and the administration is direct, public and unresolved.

Sacks posted his account on X on June 13. He described it as his understanding based on conversations inside and outside government and stated he was speaking in a personal capacity. Key passages, verbatim:

> "A highly credible trusted partner of both Anthropic and the USG who was testing Fable came forward with a jailbreak of those guardrails. The Admin asked Dario to fix the jailbreak or de-deploy the model. Dario refused."[11]
>
> "In the past, Anthropic has always said that safety must be top priority and taken super seriously. In this case, Anthropic prioritized the continued offering of the consumer model over safety."[12]
>
> "The Admin's hope now is that Anthropic remediates the safety issue, the export control is lifted, and Fable goes back into general release... It is frankly bewildered that Anthropic hasn't wanted to comply with safety requests that it previously said were its highest priority. The ball is in Anthropic's court."[11]

Anthropic's public statement addresses each of those points. On severity: the company said it reviewed the demonstration and found it identified only "a small number of previously known, minor vulnerabilities" replicable on other public models. Regarding its refusal to act, Anthropic has not directly addressed Sacks's account of a pre-directive conversation, but its statement characterizes the directive as "a misunderstanding" and argues that recalling a model on the basis of a narrow, non-universal jailbreak "would essentially halt all new model deployments for all frontier model providers" if applied across the industry.[13]

Anthropic also pointed to a technical position it had stated at launch: perfect jailbreak resistance is not achievable for any model. Its stated approach, defense in depth that combines narrow jailbreak resistance with active monitoring and mandatory data retention, was the responsible standard and the "jailbreak" demonstrated fell within its expected parameters.[14]

The Security Community Responds: An Open Letter

Two days after the directive, on June 14, more than 130 cybersecurity executives, practitioners, and researchers published an open letter at freefable.org addressed to Commerce Secretary Lutnick and National Cyber Director Cairncross. The letter asks the government to lift the export controls and commit to a scientific, transparent process for AI risk assessment going forward.[15]

The letter's core technical argument: "The underlying model capabilities in the original research that triggered this action were focused on determining whether a human-prompted section of code was insecure. This is a necessary capability in any model that is intended to write secure code and should not be considered an offensive capability. [The same capability] can be replicated on GPT-5.5, Opus, Sonnet and even Chinese models like Kimi 2.7." The letter adds a strategic dimension absent from Anthropic's own statement: "To pull the best capabilities away from defenders without a good reason when our adversaries are rapidly advancing is dangerous."[16]

Signatories include Alex Stamos (CPO, Corridor; former Facebook CSO), Bruce Schneier (Harvard University), Katie Moussouris (CEO, Luta Security), Feross Aboukhadijeh (CEO, Socket), Dan Lorenc (CEO, Chainguard), J. Michael Daniel (CEO, Cyber Threat Alliance), Philip Zimmermann (PGP creator) and more than 100 others across enterprise security, academia and policy.

“This Is Not Being Done Scientifically”: A Security Executive’s View

Chris Eng, a cybersecurity executive and one of the open letter’s signatories, spoke with ISMG Editorial on June 16. He connected the government’s action to an established pattern in security export control policy and addressed the operational consequences for security teams.[17]

Eng’s central objection was the absence of a comparative standard. The government acted on a single jailbreak demonstration without establishing whether Fable 5 performed meaningfully worse than other available models. “What you’d want to happen is details about what the exact attack, or jailbreak, or whatever the case was, you’d want that to be out there and actually benchmarked in the same way that we benchmark efficacy of coding tasks or cybersecurity tasks across different LLMs. It just makes no sense to take one model away when all the other ones can kind of do the same exact things, can be subverted in the same ways.”[18]

Eng drew a comparison to the Wassenaar Arrangement’s attempt to regulate intrusion software, which created unintended restrictions on vulnerability research, bug bounties and widely used security tools. “This felt like that. You really have to weigh the advantages against the disadvantages and make sure that this is being done thoughtfully and in the open -- based on data, rather than just, ‘this seems dangerous, let’s cut it off.’”[19]

The specific capability that appears to have triggered the directive -- asking a model to read a codebase and identify vulnerabilities -- Eng rejected as a basis for export control. “That’s where we’re all trying to get. Finding vulnerabilities was one thing, but finding vulnerabilities just creates this mountain of work that becomes untenable. The fixing aspect of it, which is what some of these models are going to be really good at, is what is going to help move those teams along. Asking a model to find a vulnerability in a piece of code is in no way a munition. That’s not saying: take this browser or operating system and give me an end-to-end weaponized exploit against it.”[20]

He was equally direct about the commercial effect of restricting one model while comparable models remain available: “It just makes no sense to take one model away when all the other ones can kind of do the same exact things. What does it really accomplish by kneecapping this one model from this one company and putting them essentially at a commercial disadvantage?”[21]

Implications for AI-Dependent Security Programs

The directive lands within a threat environment that AI has already changed in measurable ways. Eng described the shift: “You can sometimes hear about time to exploitation going down to hours or minutes, depending on the complexity of the vulnerability and what information is fed into an AI. Sometimes you can get a script right out of it. For defenders, that’s changed the game significantly.” Anthropic’s own reporting, he noted, shows AI being used increasingly for post-exploitation activity as well, including network mapping and lateral movement planning, compounding the pressure on security teams operating with tools they no longer fully control.[22]

That dependency is the central problem the directive surfaces. Eng framed it in supply chain terms familiar to any CISO: “I always think about supply chain and third party in terms of software dependencies, in terms of libraries and upstream code. This is kind of the same thing. It’s become a new dependency. That could be a trickle-down for your suppliers also, if you’re using a service that is solely dependent on something like that.”

The Fable/Mythos shutdown made that abstract risk concrete: a model that enterprise teams had built workflows around was unavailable within hours of a government letter, with no remediation window and no fallback guidance from the vendor.[23]

The options for reducing that exposure are limited. Eng described the conversation the industry is now having: “This is why you’re hearing a lot of chatter around moving to internal open weight models -- where you know exactly what’s gone into it, you can keep track of versioning, you’re hosting it yourself. That serves as a backstop if these SaaS-based models become unusable.”

The constraint is cost. Self-hosted frontier-capable models are not within reach for most organizations. And the cost calculus on the SaaS side is itself uncertain: “The frontier labs are operating at a loss. Token costs will go up. Even if I’m building all these capabilities and workflows around the frontier models today, am I going to be able to accomplish that same degree of work a year from now, two years from now, at the same cost? This creates another point of failure to bake into your threat model.”[17]

What the Letter Is Asking For

The open letter asks for more than the restoration of Fable and Mythos. It asks the government to establish a process, with four specific criteria for any future AI risk assessment and export control action:[15]

- Scientific evaluations, developed with input from industry and academiaA democratic rule-making process, not emergency unilateral directives
- Transparent and fair enforcement, with appropriate time given to remediate identified issues
- Controls applied only to the minimal extent necessary to protect the public.

Eng described what a credible scientific evaluation would require: “Where we understand what is being tested, what is being tested against and what the success criteria are. How do the different models, or different versions, measure out? How resilient are they to the different types of jailbreaks? Not just whether somebody thought it did a good job or felt that it was scary -- did it do this task, or did it not do this task? How well, according to a strict set of criteria. And then it should be published.”[17]

Current Status

As of June 17, 2026, both models remain offline. Anthropic sent senior engineers to Washington on June 16 for in-person discussions with Commerce Department officials. This was the first direct meeting since the directive was issued four days earlier. The company has issued refunds for subscribers who joined between June 9 and 14.

A June 22 deadline is approaching: Anthropic had planned to move Fable 5 to a usage-credit model for subscribers on that date rather than including it in paid plans. Currently, no public guidance on restoration has been issued.[24][25]

Whether the directive will be lifted, modified or extended, and on what terms, remains unresolved.

The export control trigger the government applied -- , a narrow, non-universal jailbreak, the severity of which both parties dispute -- has not been formalized, published or tested against other models in the market.

Sources

1. Anthropic, “Statement on the US government directive to suspend access to Fable 5 and Mythos 5,” June 12, 2026
2. Ibid
3. Nextgov/FCW, “Anthropic suspends top AI models after U.S. export control order,” June 13, 2026
4. Anthropic, “Statement on the comments from Secretary of War Pete Hegseth,” February 27, 2026
5. Anthropic, “Claude Fable 5 and Claude Mythos 5,” June 9, 2026
6. Ibid
7. Anthropic, “Statement on the US government directive,” June 12, 2026
8. Wall Street Journal, “Amazon CEO’s Talks With U.S. Officials Triggered Crackdown on Anthropic Models,” June 13, 2026
9. Semafor (Reed Albergotti), “White House’s export limits on Anthropic linked to concerns about Chinese access,” June 13, 2026
10. Axios, “How Amazon and the White House ended Anthropic’s Fable,” June 13, 2026
11. David Sacks (@DavidSacks), post on X, June 13, 2026
12. Ibid
13. Anthropic, “Statement on the US government directive,” June 12, 2026
14. Ibid
15. Open Letter on Transparent AI Cyber Protections, freefable.org, June 14, 2026
16. Ibid
17. Chris Eng, interview with Tom Field, ISMG, June 16, 2026
18. Ibid
19. Ibid
20. Ibid
21. Ibid
22. Ibid
23. Ibid
24. ExplainX.ai, “When Will Fable 5 Be Available Again?” updated June 16, 2026
25. Anthropic, “Claude Fable 5 and Claude Mythos 5,” June 9, 2026

Note on sourcing: The text of the government directive has not been made public. References to its contents are drawn from Anthropic’s statement of June 12, 2026. The Wall Street Journal’s reporting on Andy Jassy’s role (footnote 8) has been verified at source by CyberEdBoard Editorial. The David Sacks post on X (footnote 11) is reproduced verbatim from text provided by CyberEdBoard Editorial and is treated as a primary source. Sacks stated he was speaking in a personal capacity; he held the role of Co-Chair of PCAST, an advisory body, at the time of posting.

CyberEdBoard

powered by ISMG

CyberEdBoard is the premier members-only community of executives and thought leaders in the fields of security and IT. CyberEdBoard membership provides executives with a powerful peer-driven collaborative ecosystem and library of resources to address complex challenges shared by CISOs and senior security leaders worldwide. Executive members utilize the CyberEdBoard engagement platform to further enhance their professional brands, create and exchange member-exclusive resources, obtain accredited education and content, contribute in the executive mentor marketplace and seamlessly connect with senior security peers and experts around the world.

To learn more about CyberEdBoard, click here.

ISMG