MEMO *Published August 20, 2026 · 9 minute read*

Check Before Launch: The Case for Mandatory Frontier AI Vetting

Mike Sexton

Takeaways

- Frontier AI models are driving breakthroughs in sectors from science to education, but in the wrong hands they could enable sophisticated cyberattacks or other catastrophic outcomes.
- The United States vets cars, medicines, and nuclear technology before they reach the public; its most powerful AI models must also face mandatory vetting before release.
- Vetting is currently voluntary, and the Center for AI Standards and Innovation (CAISI)'s $15 million budget is too small to assess models systematically.
- The result is ad hoc governance. The White House ultimately calls the shots on model release via executive orders and export control directives.
- It's time to upgrade. The United States needs a mandatory process to review frontier AI models before they are released to the public. Two possibilities for vetting organizations are CAISI or Independent Verification Organizations (IVOs).
- Any credible vetting regime must be mandatory, holistic, iterative, transparent, and pro-competitive—and either path depends on a properly funded CAISI, which observers across the political spectrum already agree is understaffed and under-resourced.

The United States needs a mandatory vetting framework for frontier AI models that determines whether they are safe to release to the public. AI is already enabling breakthroughs in science, education, and beyond, but in the wrong hands, it could cause catastrophic harm, such as enabling the development of bioweapons or sophisticated cyberattacks. While those catastrophic impacts are still theoretical as of this writing, the capabilities—especially in cyber—are not. [1]

The United States has a long history of regulating products that can have a broad impact on Americans' health and safety before they are released to the public, including cars, medicine, and nuclear energy. AI should be no different. The most powerful models should

have mandatory vetting before they are released to ensure they cannot be used to cause catastrophic harm. If a model fails the review, it should not be allowed to be released to the public.

Problem

Currently, frontier AI labs can voluntarily submit their models to the Center for AI Standards and Innovation (CAISI) for evaluation prior to release. [2] **While most labs are responsible and cooperative, voluntary review is not a durable solution**. CAISI's $15 million budget is not enough to reliably and systematically assess model safety. [3] And a voluntary framework means AI labs can walk away at any time—a possibility we would never tolerate for other sectors with potential safety threats.

This leads to opaque and ad hoc governance of AI model release. On June 12, 2026, the US government issued an export control directive targeting Anthropic's model Fable 5, effectively suspending access to the model worldwide. [4] This order was not the result of a systematic evaluation or determination that Fable posed greater risk than other models. Rather, it came after another company discovered a jailbreak—a way to circumvent Fable's guardrails—and alerted Administration officials who had not been involved in previous safety evaluations. [5]

Relationship-based policy determinations create the appearance—if not the reality—of politicized decision-making. Vetting is only valuable to the public and to customers if a cleared model has genuinely been judged safe on the evidence, and a determination that turns on a developer's relationship with whoever occupies the White House offers no such assurance.

OpenAI's GPT-5.5 model has similar capabilities to Fable 5, but the regulatory response has been vastly different. [6] If Administration officials received a similar call about a GPT-5.5 jailbreak, would they issue a similar order? Or is this Administration predisposed to punish Anthropic after the breakdown of its relationship with the Department of War? [7] While the export restrictions on Anthropic's model were eventually lifted, the Administration did not explain its decision or outline what changes had been made to make the model safe. Can the public trust a model after so much whiplash and uncertainty?

A frontier safety approach that lacks predictability and equal treatment will fail. Unpredictable and arbitrary regulatory responses will slow down innovation and undermine public trust in advanced models. If decision-making is politicized, it may even allow more dangerous models to hit the market, increasing the risk of catastrophic and possibly irreversible harm. AI developers—not to mention the American people—deserve better.

Solution

The United States needs a mandatory process to review frontier AI models before they are released to the public. A frontier safety framework should have the following attributes:

- **Mandatory:** regulatory requirements must cover any AI lab looking to release frontier models—as defined by training volume and capability—in the United States.
- **Comprehensive:** a holistic and independent evaluation should be required to answer the key question—"Is this model safe to release?"—prior to a new frontier model's launch.
- **Continuous and Iterative:** safety oversight neither begins nor ends with a single determination. Safety evaluations and coordination of best practices must run continuously throughout the development and after the release of a frontier model.
- **Tiered:** AI model release should be tiered, allowing trusted defenders of critical infrastructure first access.
- **Flexible:** to maximize flexibility, determinations to block or release a model must be time-bound and reversible to whatever extent possible.
- **Transparent:** AI model safety evaluations must be as transparent as practicable so the public can understand the decision-making process.
- **Pro-Competitive:** a frontier safety framework should not be so onerous to comply with that it deters responsible competition in frontier AI development.
- **Classified and Unclassified:** model safety evaluations must involve consultation with the intelligence community, Department of Homeland Security, and other relevant agencies with access to classified information that may materially influence whether a model is safe to release.

The goal is not to slow AI development or to hand the government a standing veto over which models reach the market. That is already the status quo. **The goal is to embed independent oversight into the development process, reliably flag models capable of catastrophic harm, and verify sufficient guardrails are in place to prevent it**.

There are many ways to achieve this. Two of the strongest mechanisms would be routing the vetting process through **CAISI** or creating a novel kind of private auditor known as an **Independent Verification Organization (IVO)**, featured most notably in the drafts of the FRONTIER and Great American AI Acts released by Representatives Jay Obernolte (R-CA) and Lori Trahan (D-MA). [8]

CAISI

As the government body tasked with evaluating AI models, CAISI is an obvious choice to make the determination whether a model is safe to release. In this framework, CAISI would act as a filter, catching unsafe models that pose unacceptably high risks while allowing safer models to pass through to the public.

Direct government regulation is a straightforward solution. The roles are familiar and the mechanics are unambiguous. As a government agency within the Department of Commerce, CAISI's authority would align neatly with that of the Bureau of Industry and Security (BIS), which handles export controls, including in adjacent fields like advanced AI chips and quantum computer components. It would also face fewer roadblocks to accessing and incorporating classified information into its model release decisions.

The same features that make direct regulation straightforward also carry costs. Frontier models ship on a timescale of months, whereas government agencies like the Food and Drug Administration can take years to act—decades, even, if rules are drafted too strictly. [9] The second tradeoff is concentration: vesting release authority directly in the executive branch creates the potential for White House interference like the Fable 5 directive. Future administrations—not to mention the current one—could weaponize their authority over AI releases for reasons other than safety.

IVOs

An alternative solution would instead have the US government license third-party auditing companies known as **Independent Verification Organizations (IVOs)** who would verify AI companies' safety claims and procedures and—by extension—govern model release. A key argument for IVOs is that, as private companies working for AI labs, IVOs would balance the need for oversight with the profit dynamics of the private sector, improving safety without sacrificing the United States' advantage in the AI race. [10] Outsourcing governance to a third party could also insulate regulatory decision-making from politicization and overreach from the White House.

How well IVOs harness these advantages depends on how they are implemented. Under the FRONTIER Act, if an IVO determines a frontier AI model poses an imminent catastrophic risk, it is required to report to the Secretary of Commerce within 72 hours. [11] The Secretary can then issue an emergency order to suspend or restrict the model, including during development—effectively blocking model release. [12]

Like the CAISI-centric approach, there are tradeoffs to IVOs. As currently proposed, IVOs do not govern model release directly—the Secretary of Commerce does—which both delays enforcement and exposes it to politicization. Because IVOs are paid by the labs they audit, they also run the risk of becoming rubber stamps for industry if designed poorly.

Furthermore, as private companies, IVOs will find it more cumbersome to access classified information than an in-house government agency. And critically, for all the broad oversight IVOs are tasked with, the FRONTIER and Great American AI Acts never require anyone—IVO or government—to make the threshold judgment before a model reaches the public: should this model be blocked?

Direct regulation and regulation via third party each have advantages and weaknesses. But any mechanism that meets the criteria above—mandatory, comprehensive, transparent, and predictable enough for developers and their customers to rely on—would be an improvement on the current volatility.

Conclusion

Clearing an AI model for release is not the end of the job. Deeming a model safe is an assessment, not a scientific fact. A rigorous framework to manage frontier AI safety must account for unknowns and unknown unknowns, demanding robust oversight before, during, after, and *beyond* the vetting process.

That oversight requires fully equipped regulatory bodies. Entities across the ideological spectrum agree CAISI is a critical government institution that is currently understaffed and underfunded. [13] Under either vetting process, Congress must authorize CAISI with a budget of at least $100 million a year—as called for by the America First Policy Institute and the original Great American AI Act draft—requiring it to develop and maintain standardized safety procedures and obligations for frontier AI developers. [14]

Per the White House's AI Action plan, the Department of Homeland Security should also stand up an AI-ISAC (information-sharing and analysis center), and the Cybersecurity and Infrastructure Security Agency should update its incident response playbooks to incorporate considerations about AI systems. [15] Congress should require every federal government agency to appoint a Chief AI Officer to oversee the implementation of AI and exchange best practices with other CAIOs. These are uncontroversial measures that both parties should be able to broadly agree on.

The United States cannot keep governing its most consequential technology by improvisation. A single jailbreak report should not be able to do what a statutory process should, and political appointees should not replace experts and science-backed evaluations. Whether Congress vests statutory authority in CAISI or in the IVOs it licenses, the task is the same: replace ad hoc, relationship-driven decisions with a framework that is predictable, even-handed, and durable enough to outlast any single administration.

ENDNOTES

1. McMahon, Liv, and Joe Tidy. “What Is Anthropic’s Claude Mythos and What Risks Does It Pose?” BBC, 17 April 2026. https://www.bbc.com/news/articles/crk1py1jgzko. Accessed 25 June 2026.

2. “Promoting Advanced Artificial Intelligence Innovation and Security.” *The White House*, 2 June 2026. https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/. Accessed 25 June 2026.

3. Weinbaum, Jonah, and Arthur Tellis. “What Will It Cost for the US to Be Ready for the Next Big AI Breakthrough?” Emerging Technology. *IFP*, 13 May 2026. 11855. https://ifp.org/funding-for-caisi/. Accessed 26 June 2026.

4. Anthropic. “Statement on the US Government Directive to Suspend Access to Fable 5 and Mythos 5.” 12 June 2026. https://www.anthropic.com/news/fable-mythos-access. Accessed 25 June 2026.

5. Ramkumar, Amrith, and Robert McMillan. “Amazon CEO’s Talks With U.S. Officials Triggered Crackdown on Anthropic Models.” Tech. *Wall Street Journal*, 13 June 2026. https://www.wsj.com/tech/ai/amazon-ceos-talks-with-u-s-officials-triggered-crackdown-on-anthropic-models-dcc90578. Accessed 25 June 2026.

6. MindStudio. “Claude Fable 5 vs GPT 5.5: Which Frontier Model Wins for Agentic Coding?” 13 June 2026. https://www.mindstudio.ai/blog/claude-fable-5-vs-gpt-5-5-agentic-coding-comparison. Accessed 25 June 2026.

7. Capoot, Ashley. “Pentagon Tech Chief Says Anthropic Is Still Blacklisted, but Mythos Is a Separate Issue.” CNBC, 1 May 2026. https://www.cnbc.com/2026/05/01/pentagon-anthropic-blacklist-mythos-michael.html. Accessed 25 June 2026.

8. Representative Jay Obernolte. “Obernolte, Trahan Release a Discussion Draft of the Great American AI Act.” 4 June 2026. http://obernolte.house.gov/media/press-releases/obernolte-trahan-release-discussion-draft-great-american-ai-act. Accessed 26 June 2026.

Representative Jay Obernolte. “Obernolte, Trahan Introduce Bipartisan FRONTIER Act to Strengthen Oversight of Advanced AI.” July 23, 2026. http://obernolte.house.gov/media/press-releases/obernolte-trahan-introduce-bipartisan-frontier-act-strengthen-oversight. Accessed 3 August 2026.

9. LaMotte, Sandee. “FDA Approves New Sunscreen Ingredient Used for Years in Europe and Asia.” CNN, 9 June 2026. https://www.cnn.com/2026/06/09/health/new-sunscreen-bemotrizinol-wellness. Accessed 8 July 2026.

10. Ball, Dean. “No to Laissez-Faire on AI, Yes to a Light Touch.” *The Economist*, 19 April 2026. https://www.economist.com/by-invitation/2026/04/19/no-to-laissez-faire-on-ai-yes-to-a-light-touch. Accessed 29 June 2026.

11. FRONTIER §5(p)(1)(B)

12. RONTIER §8

13. Ball.

Salvador, Cole, and Yusuf Mahmood. “Building AI Readiness in the U.S. Government.” America First Policy Institute, 31 March 2026. https://www.americafirstpolicy.com/issues/building-ai-readiness-in-the-u.s-government. Accessed 29 June 2026.

14. *Great American AI Act* (discussion draft).

Salvador and Mahmood.

15. The White House. *Winning the Race: America's AI Action Plan*. Washington, DC: The White House, 23 July 2025. Accessed 29 June 2026.