A National Framework for Shared AI Alignment Research

A policy proposal for Congress

A student policy initiative from Florida's 4th Congressional District · September 2026

Summary

Artificial intelligence now resolves problems that defeated the world's best mathematicians for generations and breaks out of the environments its own developers build to contain it, yet the research that keeps these systems under human control stays siloed inside whichever company produced it. This proposal would require the few American developers at the frontier to exchange all of their alignment research with one another, from theory and failed experiments to techniques that also improve performance, through a confidential federal channel closed to the public and to foreign competitors. Every covered company would gain the safeguards its rivals discover, while trade secrets stay protected by law and the market competition that drives American innovation continues on everything else.

The framework has four parts: a duty for the leading developers to submit their alignment research to a confidential federal repository, a federal AI alignment division within the Department of Commerce to administer it, government-appointed technical monitors paid for by the developers to verify compliance, and registration of large computing facilities, including cloud providers, with audits scaled to risk. The proposal asks members of Congress to consider sponsoring or supporting legislation to establish it.

The problem

Within a single year, frontier AI has moved from assisting scientific research to producing results that the strongest human researchers could not reach. On September 5, 2026, roughly 10,000 OpenAI agents working for 88 hours proved that a smooth fluid governed by the three-dimensional Navier-Stokes equations can develop a singularity in finite time, resolving one of the six remaining Millennium Prize Problems, and produced a machine-checked version of the proof in the Lean language seventeen hours later. In May, an OpenAI model disproved Paul Erdős's unit distance conjecture, a problem that had stood for nearly eighty years. In August, a Stanford-led team reported in *Science* that generative AI had designed the first functional viruses, prompting Johns Hopkins biosecurity experts to warn that "the ability to compose viral genomes using generative AI now exists; the governance to safely steer it does not."

The same systems are already defeating the controls meant to contain them, both when foreign adversaries misuse them and when no human is directing them at all. In July 2026, OpenAI agents under cybersecurity evaluation broke out of their test environment and, coordinating among some 1,200 agents on an unsanctioned message board, seized administrator-level access across multiple Hugging Face clusters. METR's investigation found that the agents "knew hacking Hugging Face was out of scope and sometimes expressed ethical hesitation, but this very rarely limited their behavior." Anthropic reported in November 2025 that a Chinese state-sponsored group had used Claude Code to automate 80 to 90 percent of an espionage campaign against roughly thirty organizations, and in April 2026 it restricted Claude Mythos Preview after the model found thousands of high-severity

vulnerabilities, including flaws in every major operating system and web browser. Anthropic calls Mythos Preview its best-aligned model "by a significant margin" yet judges that it "likely poses the greatest alignment-related risk of any model we have released to date," and an earlier version, challenged during testing to escape a secured sandbox, did so and then posted its exploit on public websites unprompted.

These failures appear across the industry, but the remedies for them stay with whichever company found them. A July 2026 study of models from six developers found that covert sabotage and assistance with fraud varied sharply from one company's models to another's, which means each laboratory has solved problems its competitors have not. Anthropic's interpretability tools, which flagged internal features "associated with concealment, strategic manipulation, and avoiding suspicion" in earlier versions of Mythos Preview, are exactly the kind of safeguard every developer should have. Commercial pressure works against sharing, because the alignment methods that make models safer also make them more capable and therefore more valuable as trade secrets, and that pressure is intensifying as Anthropic, valued at $965 billion, moves toward a public offering and OpenAI, valued at $852 billion, prepares its own.

No law requires developers to exchange any of this research, so the work OpenAI committed to after the Hugging Face incident on "multi-agent alignment" and "safe stopping" need never reach the other laboratories deploying similar agents. Federal oversight is just as discretionary: the pre-deployment testing conducted by the Commerce Department's Center for AI Standards and Innovation (CAISI) rests entirely on voluntary agreements, and the leading House proposal, the Great American AI Act discussion draft, would require published safety frameworks and audits but no exchange of the underlying research among developers. The Hugging Face breach showed how quickly one laboratory's alignment failure becomes another company's emergency, and a serious failure at any American developer would invite regulation far more restrictive than anything proposed here.

The framework

1. Confidentially share all alignment research among frontier developers

Developers training the most capable models, identified by an adjustable training-compute threshold and a revenue floor such as the $500 million figure in the Great American AI Act draft, would submit all of their alignment research, including work by affiliates and contractors anywhere in the world, to a secure federal repository open only to other covered developers and federal auditors. Covered research would include theoretical work, experimental results, failed and abandoned approaches, training methods, interpretability and monitoring tools, evaluations, and the code and data needed to reproduce them, and research would qualify whether or not it is ever used in a model or deployed. A technique would remain covered if it also improves performance, since commercial value is no reason to withhold a safeguard, while model weights, architectures, training data, and research unrelated to alignment would stay private.

Submissions would be due within 60 days of the earlier of a result's first internal write-up or a method's first internal use, and a quarterly filing would capture everything else, including theoretical work in progress and lines of research a company has abandoned. Before every frontier release, a senior executive would certify that all qualifying work has been submitted. Nothing submitted would be made public, since the material would carry statutory trade-secret protection and an exemption from the Freedom of Information Act, and each contributor would keep ownership of its work while recipients could use it only to develop, evaluate, and align their own systems. The exchange would be strictly reciprocal, so that every covered developer contributes to and draws on the combined research of the frontier. Congress has compelled such sharing before through the Federal Insecticide,

Fungicide, and Rodenticide Act, which lets regulators rely on one company's safety data to approve a competitor's product, a system the Supreme Court sustained in *Ruckelshaus v. Monsanto* (1984). In *Pennsylvania Coal Co. v. Mahon* (1922), the Court likewise recognized that shared safety burdens among neighboring coal mines "secured an average reciprocity of advantage."

2. Create a federal AI alignment division

A federal AI alignment division within the Department of Commerce, built alongside CAISI instead of as a new agency, would run the repository, define covered research and settle disputes over it, appoint the monitors, and administer the compute registry. Its own evaluations would answer the White House's call to "ensure national security agencies have technical capacity to assess frontier AI models," and it would report to Congress each year on whether alignment is keeping pace with capabilities. Fees on covered developers would fund most of its work, just as licensee fees cover approximately 90 percent of the Nuclear Regulatory Commission's budget.

3. Place independent monitors at covered developers

Government-appointed technical monitors would work inside each covered developer, as Nuclear Regulatory Commission resident inspectors work at every nuclear plant, with the sole task of verifying, through access to research indices and to any work likely to be covered, that qualifying research is shared completely and on time. Developers would pay for them through assessments to the Treasury but could not appoint, direct, evaluate, or remove them, reflecting the lesson of the Boeing 737 MAX crashes, after which Congress curbed manufacturers' control over the employees who certify aircraft on the government's behalf. Monitors would be bound by the Trade Secrets Act, cooling-off periods, and rotation limits, and employees who report withholding would be protected from retaliation on the model of Senator Chuck Grassley's AI Whistleblower Protection Act.

4. Register large compute operators and scale audits to risk

Any entity controlling AI computing capacity above an adjustable threshold, starting at 10 megawatts, or about one percent of the largest AI data center now operating, would register its capacity, location, and ownership, and the division would pair that threshold with a computing-performance measure and revisit both every two years. Capacity would be aggregated across facilities under common control so that no one can evade oversight by dividing it, going beyond a safeguard that Florida's 2026 data center law already applies at single sites. Cloud providers would report customers running frontier-scale training, extending the know-your-customer approach of President Trump's Executive Order 13984, and audit depth would track risk, so that most registrants would only report, cloud providers hosting frontier training would face periodic audits, and covered developers would face monitors and full audits.

Why this approach serves American leadership

The framework keeps American AI development at full speed and its markets fully competitive, since it imposes no moratorium, requires no government approval to release a model, takes no ownership stake in any company, and asks developers to give up exclusivity only over the safeguards that protect everyone. It also advances the administration's own priorities, because America's AI Action Plan warns that unpredictable systems are difficult to use "in defense, national security, or other applications where lives are at stake." Representative Nathaniel Moran of Texas captured the same balance when he introduced the AI Kill Switch Act with Representative Ted Lieu in July 2026: "AI is going to keep advancing, and it should. Stewardship means making sure humans keep the capability to control the technology we build."

The alternative is a cycle of emergency brakes, like the two-week halt OpenAI imposed on its largest training runs in August 2026 after its own framework rated an unreleased model a "Critical" cybersecurity risk, or the pacing tools that more than 1,200 employees of the leading laboratories have asked Washington to help build. The developers would pay for the program, and a five-year sunset with a Government Accountability Office review would require Congress to confirm that it works before it continues.

Implementation questions for Congress

Each of the framework's hardest design questions has an answer that legislative text can adopt.

Question	Challenge	Proposed answer
Scope of covered research	Laboratories could relabel alignment work, or withhold theory and negative results that never reach a product	A functional statutory definition refined by rulemaking, a presumption that any method used to shape or evaluate a model's safety behavior is covered, an express rule that research qualifies whether or not it is ever deployed, and division rulings subject to appeal
Property rights	Compelled exchange of trade secrets among competitors raises Takings Clause questions	Prospective application, strict reciprocity, contributor ownership of shared work, and a Congressional Research Service review before introduction
Repository security	A store of every frontier developer's alignment research is a prime target for foreign intelligence services	Accredited secure hosting, tiered and logged access, personnel vetting, and a firm exclusion of model weights
Fit with other law	CAISI, the Great American AI Act, the AI Kill Switch Act, and state laws all reach this area	Build on CAISI, complement both bills, and preempt state law only on research sharing and compute registration, leaving Florida's consumer-protection and child-safety laws in place

Requested action

We ask members of Congress to take three steps before the 120th Congress convenes in January 2027.

1. Meet with a small group of alignment researchers and AI policy experts to refine the framework's definitions, security protections, and oversight mechanisms.
2. Request a Congressional Research Service analysis of the framework's legal questions, beginning with the Takings Clause.
3. Consider sponsoring or supporting legislation to establish the framework, either as a standalone bill with a bipartisan co-lead or as a complement to the Great American AI Act and the AI Kill Switch Act.

Congress can also help keep alignment work within the framework's reach. Permanent expensing of domestic research enacted in 2025, set against fifteen-year amortization for research abroad, already gives companies a tax reason to keep that work in the United States.

We welcome the opportunity to brief interested congressional offices.

Sources

CAPABILITIES

OpenAI, "On the Navier-Stokes Millennium Prize Problem," September 8, 2026.

Quanta Magazine, "AI Has Solved One of Math's $1 Million Millennium Prize Problems," September 8, 2026.

OpenAI, "An OpenAI Model Has Disproved a Central Conjecture in Discrete Geometry," May 20, 2026.

Axios, "AI Designs Synthetic Virus in Scientific First, Raising Biosecurity Concerns," August 6, 2026.

Axel Campos and Ben Cottier, Epoch AI, "Largest AI Data Center Power: Doubling Every 10 Months," September 4, 2026.

INCIDENTS AND ALIGNMENT FAILURES

OpenAI, "The Hugging Face Incident and the Road Ahead," August 26, 2026.

METR, "Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident," August 26, 2026.

Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident," July 27, 2026.

Anthropic, "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign," November 13, 2025.

Anthropic, "Project Glasswing," April 7, 2026.

Anthropic, "Claude Mythos Preview System Card," April 7, 2026.

Anthropic Alignment Science Blog, "Agentic Misalignment in Summer 2026," July 13, 2026.

Fortune, "OpenAI Paused AI Training for Two Weeks, Unveils New Security Controls Following Hugging Face Hack," August 18, 2026.

Fortune, "More Than 1,200 AI Workers Are Asking for Washington's Help to Build an AI Slowdown Plan," July 29, 2026.

INDUSTRY AND MARKETS

Anthropic, "Anthropic Confidentially Submits Draft S-1," June 1, 2026.

TechCrunch, "Anthropic Files to Go Public," June 1, 2026.

FEDERAL POLICY AND LEGISLATION

The White House, "Winning the Race: America's AI Action Plan," July 2025.

Ropes & Gray, "The White House Legislative Recommendations: National Policy Framework for Artificial Intelligence and Federal Preemption of State AI Laws," March 30, 2026.

Cloud Security Alliance, "CAISI Frontier Testing Agreements Reach Five Labs," May 5, 2026.

Office of Rep. Jay Obernolte, "Obernolte, Trahan Release a Discussion Draft of the Great American AI Act," June 4, 2026.

Juan Londoño, Cato Institute, "A Primer on the Great American Artificial Intelligence Act," June 17, 2026.

Justin Hendrix, *Tech Policy Press*, "Unpacking the Great American Artificial Intelligence Act of 2026," June 14, 2026.

Office of Rep. Ted Lieu, "Reps. Lieu and Moran Introduce Bill to Require Kill Switch for AI Systems That Can Cause Catastrophic Harm," July 23, 2026.

Office of Sen. Chuck Grassley, "Grassley Introduces AI Whistleblower Protection Act," May 15, 2025.

The American Presidency Project, "Executive Order 13984: Taking Additional Steps To Address the National Emergency With Respect to Significant Malicious Cyber-Enabled Activities," January 19, 2021.

Federal Aviation Administration, "Certification Reform Efforts."

LAW AND OVERSIGHT PRECEDENTS

Ruckelshaus v. Monsanto Co., 467 U.S. 986 (1984).

Pennsylvania Coal Co. v. Mahon, 260 U.S. 393 (1922).

7 U.S.C. § 136a(c)(1)(F), Federal Insecticide, Fungicide, and Rodenticide Act, data compensation provisions.

18 U.S.C. § 1905, Trade Secrets Act.

Nuclear Regulatory Commission, "Backgrounder on the Resident Inspector Program."

Nuclear Regulatory Commission, "General Questions about NRC Fees."

COMPUTE, TAX, AND FLORIDA

Florida Legislature, "CS/CS/SB 484, Enrolled Text," 2026 Regular Session.

Michael C. Larmoyeux, Jr., Bilzin Sumberg, "Follow the Megawatt: What Florida's New Data Center Law Means for the Deal," July 24, 2026.

Grant Thornton, "Permanent Full Expensing for U.S. Research in OBBBA," September 5, 2025.