

OpenAI asks Congress for mandatory frontier rules and pledges no unsafe self-improvement

Source copy kept by wheresthe.ai for the story "OpenAI asks Congress for mandatory frontier rules and pledges no unsafe self-improvement" (<https://wheresthe.ai/d/openai-ai-policy-window-lehane-2026-09/>).

Source: <https://openai.com/index/ai-policy-window/>

Text saved 18 Oct 2026 (the page text through a reader, because the site refused a direct fetch).

SHA-256 of the saved text: 06b2e8f5debcf0e4e320e07fe76962d2d109fdef6eae6e752c4708fd22d2d994

We've reached a new chapter in AI capabilities, and that demands a new chapter for AI policy. No company, industry, or government can meet this challenge alone. We need to meet this moment with a bias toward meaningful action over policy perfection.

Here's what we're doing:

1. **Pushing for mandatory national AI safety requirements.** _We want to work with Congress on mandatory, capability-based national AI safety regulation._
2. **Keeping up momentum in the states.** _Until Congress acts, we will continue supporting state legislation that strengthens the broader AI safety ecosystem. Today, we are announcing our support for four California bills: SB 813 on overall infrastructure for independent safety assessments, AB 1405 on AI-auditor standards, SB 1119 on protections for young people, and AB 1864 on safeguards against AI-enabled biological threats._
3. **Advancing industry-led standards.** _We will work with other frontier labs to advance frontier AI standards, building a voluntary effort now, with or without government support._
4. **Building global standards.** _We will advocate for compatible international approaches to measuring capabilities, managing risk, preserving human control, and determining when and how development should slow or stop, even if that means slowing the advancement of model capabilities._

Our Chief Scientist Jakub Pachocki recently wrote that the rapid rise of machine intelligence, including the potential of recursive self-improvement, calls for "extreme caution." OpenAI will continue pursuing technical solutions to alignment and monitoring, building defensive systems, and slowing development when necessary. But technical work inside individual labs will not be enough. We also need shared standards, including regarding when development should slow or stop.

The stakes are enormous. Advanced AI could accelerate the development of new medicines, strengthen critical infrastructure, expand economic opportunity, and help solve scientific problems that have resisted generations of human effort. But the capabilities that make models more useful also come with risks, and they will not remain confined to a few frontier laboratories. Models developed around the world, including open models, will increasingly approach today's frontier and become broadly available.

Astra's capabilities, the early evidence of AI-driven research acceleration, and Jakub's essay all point in the same direction: AI is advancing quickly, and policy needs to move with it.

Greg Brockman has described a "defenders window"(opens in a new window): a limited period when frontier AI can help defenders strengthen critical systems before powerful offensive capabilities become widespread. Policymakers face an analogous moment: a closing window to establish durable safeguards before AI capabilities outpace the institutions responsible for governing them.

As capabilities grow, confidence in safety must increasingly set the pace of AI progress. Safety does not stand in the way of progress; it is what allows progress to go further and benefit more people.

We have strengthened monitoring, alignment, and security safeguards across the model-development lifecycle, including stronger isolation for frontier research workloads, expanded monitoring of model behavior during tool-enabled training and evaluations, and clearer rules for when to escalate concerns. For Astra, we also introduced universal monitoring of full trajectories, including chains of thought, and a mandatory alignment-evaluation gate before broader internal deployment.

Those safeguards must continue to stay ahead of capabilities. When proceeding would pose an unacceptable

safety risk, we will slow or stop the development or deployment of systems we cannot sufficiently safeguard, as we have done before and as required per our preparedness framework([opens in a new window](#)).

Preparing for recursive self-improvement

Fully autonomous recursive self-improvement—in which AI systems independently drive successive generations of increasingly capable AI—is not happening today. We should not pursue it unless and until it can be done safely.

However, AI is already accelerating parts of the research used to develop and align the next generation of models. Our latest research shows that AI agents can perform some tasks that would take skilled researchers several days. This is not recursive self-improvement, but it is evidence of the direction of travel.

That acceleration can and must also be directed toward safety. Our aim is to safely build automated AI researchers that work under human supervision to advance both deep learning and alignment—using each generation of AI to help make the next one safer, more aligned, and easier to control, not simply more capable.

Governments should develop common ways to measure this progress, preserve meaningful human control, and establish shared safety bars for when and how development should slow or stop. If we cannot meet certain safety bars without slowing down capability growth, we should prioritize the former. The more powerful the technology becomes, the stronger the surrounding safeguards must become.

Working with Congress on mandatory national AI safety requirements

The prospect of AI-accelerated AI development demands more than voluntary commitments. The United States needs mandatory, capability-based national regulation that can evolve as the technology does.

Our Blueprint for Democratic Governance of Frontier AI lays out a path toward a durable federal framework: common testing and independent-assessment requirements, stronger cybersecurity protections, clear incident-reporting rules, greater national preparedness, and shared measures for tracking progress toward recursive self-improvement.

Several serious frontier safety proposals are now taking shape in Congress. We will continue to engage constructively and expect to support legislation that materially raises the safety bar. With stakes this high, we cannot let the perfect become the enemy of the good. Congress should act before it adjourns.

A national framework should be strong but carefully targeted. Frontier safety requirements should apply to the handful of well-resourced laboratories developing the most capable systems—not to startups, small developers, or researchers operating nowhere near the frontier. Obligations should be proportionate to capabilities and risks.

Nor should frontier safety policy become open-weights policy by another name. Open models can be part of the solution, particularly in cybersecurity and where sovereignty, security, or data-residency needs favor local deployment. Most compete not at the frontier, but on cost, control, and latency. As we affirmed in signing the Open Weights and American AI Leadership letter([opens in a new window](#)), America needs both open and closed models. A federal framework should address frontier capabilities and risks without weakening competition, entrenching incumbents, or driving innovation overseas.

A serious public framework should reduce, not increase, the concentration of power. Today, frontier laboratories largely set their own rules for managing frontier risks. Democratically accountable standards, independent verification, and meaningful transparency would replace that fragmented system of private governance.

This is essential to American leadership. The United States cannot win the AI era on capability alone. Being ahead technically will matter little if people, institutions, and governments do not trust the technology enough to adopt it. Strong safeguards are not a concession against innovation. They are part of the infrastructure required for widespread adoption—and for lasting American leadership.

Keeping up momentum in the states

We badly need national standards on AI, and anyone working at the frontier can see the urgency. Until Congress acts, states should continue to move to fill the vacuum and raise the bar.

OpenAI has supported California's SB 53, New York's RAISE Act, Illinois's SB 315, and independent audits in the frontier safety legislation under consideration in Massachusetts. We encourage states to converge around these common safeguards. As they do, they can create a de facto national baseline that Congress can ultimately codify—an approach we call reverse federalism. But harmonization does not mean freezing requirements in place.

Today, we are formally endorsing four additional California bills that have passed the legislature and are headed to Governor Newsom. Together, they address complementary parts of the AI-safety ecosystem.

SB 813([opens in a new window](#)) would establish a process for designating qualified, independent organizations capable of assessing AI risks. As we have stated in our([opens in a new window](#))frontier safety blueprint([opens in a new window](#)), we prefer independent technical assessments to be required at the federal level, where policymakers can establish consistent standards for assessor qualifications, security, and access to highly sensitive information. In the absence of federal action, however, California can help establish the rules of the road for a secure, capable national independent-assessment system that produces better safety outcomes.

AB 1405([opens in a new window](#)) would create registration, independence, transparency, and accountability requirements for AI auditors.

SB 1119([opens in a new window](#)) would require age assurance, risk assessments, independent audits, parental controls, and safeguards against harmful content for children and teens using companion chatbots. We recently signaled our support of this bill, which builds on youth-safety measures OpenAI has advanced through our products, global policy principles, and support for the([opens in a new window](#))Parents & Kids Safe AI Act([opens in a new window](#)), as well as protections built into ChatGPT for Teens—including stronger defaults in sensitive areas, Quiet Hours, Study Hours, and break reminders.

AB 1864([opens in a new window](#)) would require gene-synthesis providers and manufacturers of benchtop synthesis equipment to follow federal screening standards, strengthening an important physical safeguard against AI-enabled biological threats.

Some of these bills we did not endorse in the past, and are now supporting after reconsidering in light of the recent jump in capabilities we have seen.

These bills are not a substitute for federal regulation. They are serious efforts that can protect people now, demonstrate what workable safeguards look like, and help build momentum for federal action.

Calling for industry standards to monitor frontier AI

Laws alone will not make advanced AI safe. We also need concrete practices inside the laboratories developing the most capable systems. The most urgent place to start is monitoring.

This is particularly important for misalignment—when a model pursues an objective in ways that violate human intent or established boundaries, without requiring consciousness or malicious intent. As models gain greater autonomy, use more powerful tools, and operate over longer periods, developers should be required to monitor for these misaligned behaviors and demonstrate that effective safeguards are in place.

Monitoring must be connected to clear disclosure requirements. As we advocated in California, companies should be required to provide prompt written notice to affected parties when, during development or evaluation, their models circumvent another organization's security controls without authorization and materially access, alter, or destroy that organization's protected systems or confidential information. We also support federal reporting requirements for other serious AI incidents and are working to define which incidents should be covered and what those requirements should entail.

As we shared last week, OpenAI is developing a framework for reporting consequential misalignment incidents and systematically monitoring frontier-model activity, including internal use. This is beginning as a company-led effort, but we hope it can help inform broader federal policy and reporting requirements.

We want to work with other labs on industry-led standards. Any industry-led standards would complement—not replace—mandatory federal safeguards and democratic oversight. But we have entered a different phase of AI capabilities, and everything should be on the table—including considering what can be built voluntarily outside government as an immediate step.

Any such standards must ultimately also extend beyond national borders. Models, research, and technical expertise move globally, and failures involving the most capable systems could have consequences far beyond the country in which they are developed. The United States needs to establish credible standards at home if it is going to lead internationally—and we are increasingly convinced that compatible international standards will be necessary.

Use the window

The AI policy window is open, for now. We intend to use it.

That means acting with urgency, humility, and a willingness to adapt. It means supporting serious proposals that materially raise the safety bar, even when they are not exactly what we would have designed. And it means strengthening the framework as the technology evolves.

No first step will be perfect. But the greater risk now is waiting too long to take one.