

Anthropic proposes laws letting governments block dangerous frontier model releases

Source copy kept by wheresthe.ai for the story "Anthropic proposes laws letting governments block dangerous frontier model releases" (<https://wheresthe.ai/d/anthropic-policy-on-ai-exponential-2026-06/>).

Source: <https://www.anthropic.com/policy-on-the-ai-exponential>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: b0e52bcebbabba8435b92f64cdff369ab4e3eaf0efff9d90c5d8e2fdc801c898

Documents the source offers: <https://www-cdn.anthropic.com/files/4zrzovbb/website/0a58d567024a8b448ff15158ebc3625328dfcc1f.pdf>

Policy on the AI Exponential AI is advancing at exponential speed, and the policymaking process was built for a slower world. We are sharing two policy proposals to prepare for AI progress. The first, our Advanced AI Framework, offers a roadmap for governing increasingly capable systems, from transparency and independent evaluation to government authority to block or deter dangerous deployments. The second, our Economic Policy Framework, turns to the question of how to prepare workers and the economy for AI's impact and ensure the financial benefits of AI are broadly shared. Together they cover two sides of the same challenge: steering the technology responsibly as it advances, and preparing society for what it brings.

0 1 Advanced AI Framework 0 2 Economic Policy Framework

Anthropic's Advanced AI Framework

Read the advanced AI Safety Framework →

We are publishing our proposal for how governments should address catastrophic risks from the most powerful AI models, including granting them the legal authority to block or deter dangerous deployments. AI capabilities are on a sharp upward trend. A few years ago, AI models could barely write code. This year, Claude Mythos Preview discovered thousands of high-severity vulnerabilities, including in every major operating system and browser. Evidence strongly suggests this trend will continue—and with it, the risk of catastrophic harm will increase, too. This calls for government action and for regulation—regulations that are carefully designed to prevent government overreach and protect innovation. When a model poses risks of this kind, the government should have the legal authority to block or deter its deployment—beyond what exists in current law or in existing proposals in Congress—with civil penalties tied to global annual revenue that escalate with repeated violations. Frontier AI developers should have to test these models, be transparent to the public about their findings, submit them to independent evaluation, and maintain a robust security program. These rules should apply only to models trained using more than 10^{25} floating-point operations (FLOPs), developed by companies earning more than \$500M in AI-related revenue or spending more than \$1 billion on AI R&D. They would address four kinds of catastrophic risk:

Biological risk. If AI systems are released without safeguards, it could become substantially easier to develop biological weapons. The same capabilities that accelerate drug discovery can be used to make it cheaper and easier for attackers to develop dangerous viruses. **Cyber risk.** Frontier AI models can now find critical software vulnerabilities at large scale. Used defensively, these capabilities can secure critical systems, but they also raise the stakes for protecting essential infrastructure like hospitals and the energy grid. **Loss of control risk.** As AI systems improve, it could become much harder to control systems that act outside of their developers' control. **Automated R&D.** AI systems are automating the research and development of AI itself, which could further amplify the three above risks. Several recent state laws require companies to describe their safety practices and share them publicly. Anthropic has supported these laws. But the rapid pace of acceleration means that transparency alone is no longer sufficient. Governments need to play a more substantial role. Our framework is written primarily with the US federal government in mind. But addressing AI risks cannot wait for action in Washington. We do not believe Congress should preempt state law unless it enacts a federal law that is at least as strong as the framework we are proposing today. Preemption is powerful, so it should be surgical, allowing states to regulate AI on all issues—such as child safety and consumer protection—which fall outside the specific safety functions a federal law covers. Our framework is available in full here . Below, we summarize its key recommendations.

Requirements for frontier developers

Transparency. Frontier developers should test their models and publish a summary of the results. They should publish a safety framework explaining how they evaluate catastrophic risks, and system cards evaluating the capabilities and risks of their frontier models. This is already required by

California and New York law. But our framework goes further by requiring that developers also publish regular risk reports that describe their overall risk posture, and regularly engage independent evaluators. Independent evaluation. Frontier developers should engage at least one qualified independent evaluator to publish a review of their evaluations and risk reports. In parallel, governments and industry should develop the ecosystem of qualified independent evaluators by setting standards for them and ensuring they have funding and sufficient access to frontier models. Security. Model weights and training infrastructure are valuable targets for cyberattackers, including well-resourced state actors. Frontier developers should secure their whole development environment and protect against threats from both outside and inside the company. They should describe their program at a high level publicly, share details with a designated agency on request, have channels to report model distillation attacks to that agency, and test their own defenses regularly. Enforcement and regulatory authority with teeth. The government should be able to block or deter the deployment of models that pose a significant risk of catastrophic harm. We must also avoid overly broad or heavy-handed regulatory power. Our framework proposes both a mechanism for blocking dangerous deployments, and concrete safeguards that would prevent that power from being misused. Policymakers could begin with a lighter-touch approach, then adapt this as model capabilities advance and the evaluation ecosystem matures. Societal resilience In the second half of our framework, we recommend ways for governments to increase society's resilience to the risks that models could pose. On biology, we recommend prevention measures such as gene synthesis screening, detection measures such as early-warning biosurveillance to spot novel outbreaks, and preparedness measures like stockpiling protective equipment and ways to suppress airborne transmission. On cyber, we recommend hardening the software that the internet runs on, deploying technical support for critical infrastructure operators, and replacing legacy software systems in critical infrastructure. For measurement and situational awareness, we recommend instituting a dedicated function within government to track frontier cyber capabilities. We also recommend government and industry co-develop model safeguards so that cyber capabilities can be shared broadly. The resilience agenda for loss-of-control and automated R&D risks is less developed. We are actively researching this, but it needs much more work across the field. Promising directions we've identified so far include developing the capacity to detect and respond to AI systems acting outside their developers' control, and building infrastructure for containing or shutting down such systems. What comes next These issues are both complex and novel, and we expect vigorous debate about the recommendations we present. But we urge policymakers to actively engage on these issues now. AI capabilities are going to improve rapidly over the coming months. Their governance needs to keep pace. You can read the full policy framework [here](#) .