EXHIBIT 179

Policy Announcements

Anthropic's Responsible Scaling Policy: Version 3.0

Feb 24, 2026

Read the Responsible Scaling Policy



We're releasing the third version of our Responsible Scaling Policy (RSP), the voluntary framework we use to mitigate catastrophic risks from AI systems.

Anthropic has now had an RSP for more than two years, and we've learned a great deal about its benefits and its shortcomings. We're therefore updating the policy to reinforce what has worked well to date, improve the policy where necessary, and implement new measures to increase the transparency and accountability of our decision-making.

You can read the new RSP in full here. In this post, we'll discuss some of the thinking behind the changes.

The original RSP and our theory of change

The RSP is our attempt to solve the problem of how to address AI risks that are not present at the time the policy is written, but which could emerge rapidly as a result of an exponentially advancing technology. When we wrote the original RSP in September 2023, large language models were essentially chat interfaces. Today they can browse the web, write and run code, use computers, and take autonomous, multi-step actions. As each of these new capabilities have emerged, so have new risks. We expect this pattern to continue.

We focused the RSP on the principle of *conditional*, or *if-then*, commitments. *If* a model exceeded certain capability levels (for example, biological science capabilities that could assist in the creation of dangerous weapons), *then* the policy stated that we should introduce a new and stricter set of safeguards (for example, against model misuse and the theft of model weights).

Each set of safeguards corresponded to an "AI Safety Level" (ASL): for example, ASL-2 referred to one set of required safeguards, whereas ASL-3 referred to a more stringent set of safeguards needed for more capable AI models.

Early ASLs (ASL-2 and ASL-3) were defined in significant detail, but it was more difficult to specify the correct safeguards for models that were still several generations away. We therefore intentionally left the later ASLs (ASL-4 and beyond) largely undefined, and hoped to develop them in more detail once we had a better picture of what higher AI capability levels would entail.

The following is a rough description of our "theory of change"—that is, the mechanisms whereby we hoped to affect the ecosystem with the RSP:

- *An internal forcing function.* Within Anthropic, we hoped the RSP would compel us to treat important safeguards as requirements for launching (and training) new models. This made the importance of these safeguards clear to the large and growing organization, spurring us on to make faster progress.
- *A race to the top.* We hoped that announcing our RSP would encourage other AI companies to introduce similar policies. This is the idea of a "race to the top" (the converse of a "race to the bottom"), in which different industry players are incentivized to improve, rather than weaken, their models' safeguards and their overall safety posture. Over time, we hoped RSPs, or similar policies, would become voluntary industry standards or go on to inform AI laws aimed at encouraging safety and transparency in AI model development.
- *Creating more consensus about risks.* We viewed the capability thresholds as potentially important moments for the industry. If we reached an important capability threshold (such as the ability of AI models to support the end-to-end production of bioweapons), we would institute the appropriate safeguards ourselves and use the evidence we'd obtained about AI capabilities to advocate to other companies and governments that they take action as well. In other words, we believed that the capability thresholds might be good points at which to go beyond *unilateral* action (Anthropic requiring safeguards for its own models) and encourage *multilateral* action (other AI companies, and/or governments also requiring such safeguards).
- *Looking to the future.* We recognized that, at some of the later capability thresholds, the intensity of countermeasures we were envisioning (for example, achieving high robustness against misuse of AI models by state-level actors) would likely be difficult or impossible for Anthropic to accomplish unilaterally. We hoped that by the time we reached these higher capabilities, the world would clearly see the dangers, and that we'd be able to coordinate with governments worldwide in implementing safeguards that are difficult for one company to achieve alone.

Assessing our theory of change

Two and a half years later, our honest assessment is that some parts of this theory of change have played out as we hoped, but others have not. The following are the areas in which the RSP has been successful:

- Our RSP **did** incentivize us to develop stronger safeguards. For example, in order to comply with our ASL-3 deployment standard (which is primarily about risks from chemical and biological weapons from threat actors with relatively modest resources and expertise), we developed increasingly sophisticated and accurate methods (specifically, input and output classifiers) to block content of concern.
- More broadly, the overall implementation of the ASL-3 standard **did** prove feasible. We activated ASL-3 safeguards for relevant models in May 2025 and have been working to improve them ever since.
- Our RSP **did** encourage other AI companies to adopt somewhat similar standards: within a few months of announcing our RSP, both OpenAI and Google DeepMind adopted broadly similar frameworks. Some companies have also implemented bioweapon-related classifiers in a similar vein to our ASL-3 defenses. The principles behind these voluntary standards, including those in the RSP, have helped to inform the development of early AI policy. We've seen governments around the world (for example in California with SB 53, in New York with the RAISE Act, and with the EU AI Act's Codes of Practice) start to require frontier AI developers to create and publish frameworks for assessing and managing catastrophic risks—requirements Anthropic addresses through public documentation including its Frontier Compliance Framework. Encouraging these kinds of rigorous transparency frameworks for the industry was exactly what our RSP had set out to do.

Nevertheless, other parts of our theory of change have not panned out as we'd hoped:

- The idea of using the RSP thresholds to create more consensus about AI risks **did not** play out in practice—although there was some of this effect. We found pre-set capability levels to be far more ambiguous than we anticipated: in some cases, model capabilities have clearly approached the RSP thresholds, but we have had substantial uncertainty about whether they have definitively passed those thresholds. The science of model evaluation isn't well-developed enough to provide dispositive answers. In such cases, we have taken a precautionary approach and implemented the relevant safeguards, but our internal uncertainty translates into a weak external case for taking multilateral action across the AI industry.
  - Biological risks provide an example of this "zone of ambiguity". Our models now show enough biological knowledge that they pass most tests we can run quickly and easily, so we can no longer make a strong argument that risks are low from a given model. But these tests alone aren't sufficient for a strong argument that risks are *high*, either. We've sought additional evidence, such as supporting an extensive wet-lab trial, but results remain ambiguous, especially because the studies take long enough that more powerful models are available by the time they're completed.
- Despite rapid advances in AI capabilities over the past three years, government action on AI safety has moved slowly. The policy environment has shifted toward prioritizing AI competitiveness and economic growth, while safety-oriented discussions have yet to gain meaningful traction at the federal level. We remain convinced that effective government engagement on AI safety is both necessary and achievable, and we aim to continue advancing a conversation grounded in evidence, national security interests, economic competitiveness, and public trust. But this is proving to be a long-term project—not something that is happening organically as AI becomes more capable or crosses certain thresholds.

As noted above, we were able to implement ASL-3 safeguards unilaterally and at reasonable costs to the operation of the company. However, this may not remain true for higher capability levels and higher ASLs. While our higher ASLs are largely undefined, the robust mitigations we laid out in the prior RSP might prove outright impossible to implement without collective action. As one illustration of the scale of the challenge, a RAND report on model weight security states that its "SL5" security standard, aimed at stopping top-priority operations by the most cyber-capable institutions, is "currently not possible" and "will likely require assistance from the national security community."

The combination of (a) the zone of ambiguity muddling the public case for risk, (b) an anti-regulatory political climate, and (c) requirements at the higher RSP levels that are very hard to meet unilaterally, creates a structural challenge for our current RSP. We could have tried to address this by defining ASL-4 and ASL-5 safeguards in ways that made compliance easy to achieve—but this would undermine the intended spirit of the RSP.

Instead, we are choosing to acknowledge these challenges transparently and restructure the RSP *before* we reach these higher levels. The revised RSP aims to adopt more realistic unilateral commitments that are difficult but still achievable in the current environment, while continuing to comprehensively map the risks we believe the full industry needs to address multilaterally.

Updating our Responsible Scaling Policy

The new version of our RSP has three key elements.

1. Separating our plans as a company from our recommendations for the industry

Our RSP now outlines two sets of mitigations: first, the mitigations that we plan to pursue regardless of what others do; and second, an ambitious capabilities-to-mitigations map that, we believe, would help adequately manage the risks from advanced AI if implemented across the AI industry.

Read the full Responsible Scaling Policy.

2. Frontier Safety Roadmap

Our new RSP introduces a requirement to develop and publish a Frontier Safety Roadmap, which will describe our concrete plans for risk mitigations across the areas of Security, Alignment, Safeguards, and Policy. Goals described in the Roadmaps are intended to be ambitious, yet achievable—providing the kind of forcing function that we consider to be a past success of our RSP.

Rather than being hard commitments, these are public goals that we will openly grade our progress towards. This strategy of "nonbinding but publicly-declared" targets borrows from the transparency approach we've been championing for frontier AI legislation (although it provides the public with much more detail than is required under existing legislation), and from the successes of our previous RSP versions.

Some example goals from our current Frontier Safety Roadmap include:

- Launch "moonshot R&D" projects to investigate ambitious, possibly unconventional ways to achieve unprecedented levels of information security;
- Develop a method for red-teaming our systems (likely involving significant automation) that surpasses the collective contributions from the hundreds of participants in our bug bounty;
- Implement a number of systematic measures to ensure Claude behaves according to its constitution;
- Establish comprehensive, centralized records of all our critical AI development activities, and use AI to analyze these records for issues including concerning behavior by insiders (both human and AI) and security threats;
- Publish a policy roadmap with concrete proposals for a "regulatory ladder"—policies that scale with increasing risk and that could help guide government AI policy.

Read the Frontier Safety Roadmap for more on these and our other goals.

3. Risk Reports and external review

Risk Reports are another way in which we're improving upon what worked well about our previous RSP. We found that producing a proto-Risk Report, our Safeguards Report from May 2025, was useful for our internal understanding and the public communication of the risks. Risk Reports extend this to a more systematic, comprehensive practice.

Risk Reports will provide detailed information on the safety profile of our models at the time of publication. They will go beyond describing model capabilities to explain how capabilities, threat models (the specific ways that models might pose threats), and active risk mitigations fit together, and provide an assessment of the overall level of risk. Risk Reports will be published online (with some redactions[1]) every 3-6 months.

The new RSP also requires external review of Risk Reports in certain circumstances. We will appoint expert third-party reviewers who are deeply familiar with AI safety research, are incentivized to be open and honest about Anthropic's safety position, and are free of major conflicts of interest. They will have unredacted or minimally-redacted access to the Risk Report and will subject our reasoning, analysis, and decision-making to a comprehensive public review. Although our current models do not yet require external review, we are already running pilots and working toward this goal.

Risk Reports will address any gaps between our current safety and security measures and our more ambitious recommendations for industry-wide safety. We are hopeful that describing and publicizing such gaps could help contribute to public awareness and thus to beneficial policy change in the future.

Read the Initial Risk Report.

Conclusion

The Responsible Scaling Policy was always planned to be a living document: a policy that had the flexibility to change as AI models become more capable. This third revision amplifies what worked about the previous RSP, commits us to more transparency about our plans and our risk considerations, and separates out our recommendations for the industry at large from what we can achieve as an individual company.

In that same spirit of pragmatism we will continue to revise and refine our RSP, and our methods of evaluating and mitigating risks, as the technology evolves.

Footnotes

1. As we discuss in the RSP, we will aim to minimize redactions to the public version of the Risk Report. Reasons we may nonetheless have to redact some of the text include legal compliance, intellectual property protection, public safety, and privacy.

Related content

Improving Fable 5's biology safeguards
Read more →

Mariano-Florentino (Tino) Cuéllar to join Anthropic as Chief Global Affairs Officer
Mariano-Florentino (Tino) Cuéllar will join Anthropic as its first Chief Global Affairs Officer.
Read more →

Investigating three real-world incidents in our cybersecurity evaluations
Read more →

Products
Claude
Claude Code
Claude Code Enterprise
Claude Cowork
@Claude
Claude Design
Claude Science
Claude Security
Claude for Chrome
Claude for Microsoft 365
Skills
Download app
Pricing
Log in to Claude

Models
Mythos
Fable
Opus
Sonnet
Haiku

Solutions
AI agents
Code modernization
Coding
Customer support
Cybersecurity
Enterprise
Financial services
Government
Healthcare
Higher education
K-12 teachers
Legal
Life sciences
Nonprofits
Small business

Claude Platform
Overview
Developer docs
Pricing
Ecosystem
Marketplace
Regional compliance
Claude on AWS
Google Cloud
Microsoft Foundry
Console login

Resources
Blog
Claude partner network
Community
Connectors
Courses
Customer stories
Engineering at Anthropic
Events
Plugins
Powered by Claude
Service partners
Tutorials
Use cases

Programs
Startups
Research Labs

Help and security
Availability
Status
Support center

Company
Anthropic
Careers
Policy
Economic Futures
Research
News
Claude's Constitution
Claude Corps
Keep thinking
Policy on the AI Exponential
Responsible Scaling Policy
Security and compliance
Transparency

Terms and policies
Privacy choices
Privacy policy
Consumer health data privacy policy
Responsible disclosure policy
Terms of service: Commercial
Terms of service: Consumer
Terms of Service: US K-12
Data Processing Agreement: US K-12
Usage policy

© 2026 Anthropic PBC