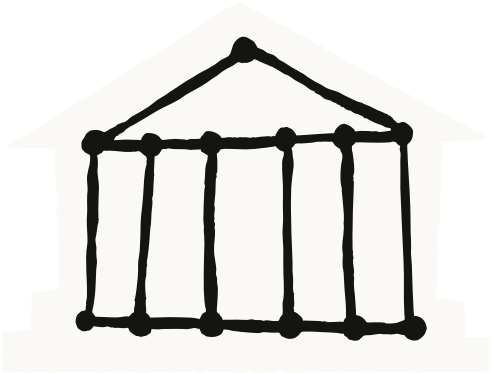
Policy

The need for transparency in Frontier AI

Jul 8, 2025 • 4 min read

Read the Transparency Framework



Frontier AI development needs greater transparency to ensure public safety and accountability for the companies developing this powerful technology. AI is advancing rapidly. While industry, governments, academia, and others work to develop agreed-upon safety standards and comprehensive evaluation methods—a process that could take months to years—we need interim steps to ensure that very powerful AI is developed securely, responsibly, and transparently.

We are therefore proposing a targeted transparency framework, one that could be applied at the federal, state, or international level, and which applies only to the largest AI systems and developers while establishing clear disclosure requirements for safety practices.

Our approach deliberately avoids being heavily prescriptive. We recognize that as the science of AI continues to evolve, any regulatory effort must remain lightweight and flexible. **It should not impede AI innovation, nor should it slow our ability to realize AI's benefits—including lifesaving drug discovery, swift delivery of public benefits, and critical national security functions.** Rigid government-imposed standards would be especially counterproductive given that evaluation methods become outdated within months due to the pace of technological change.

Minimum Standards for AI Transparency

Below are the core tenets we believe should guide AI transparency policy:

- **Limit Application to the Largest Model Developers:** AI transparency should apply only to the largest frontier model developers that are building the most capable models - where frontier models are distinguished by a combination of thresholds for computing power, computing cost, evaluation performance, annual revenue and R&D. To avoid burdening the startup ecosystem and small developers with models at low risk to national security or for causing catastrophic harm, the framework should include appropriate exemptions for smaller developers. We welcome input from the start-up community on what those thresholds should be. Internally, we've discussed the following examples for what the threshold could look like: annual revenue cutoff amounts on the order of $100 million; or R&D or capital expenditures on the order of $1 billion annually. These scoping thresholds should be periodically reviewed as the technology and industry landscape evolves.
- **Create a Secure Development Framework:** Require covered frontier model developers to have a Secure Development Framework that lays out how they will assess and mitigate unreasonable risk in a model. Those risks must include the creation of chemical, biological, radiological and nuclear harms, as well as harms caused by misaligned model autonomy. Secure Development Frameworks are still an evolving safety tool, so any proposal should strive for flexibility.

- **Make the Secure Development Framework Public:** The Secure Development Framework should be disclosed to the public, subject to reasonable redaction protections for sensitive information, on a public-facing website registered to and maintained by the AI company. This will enable researchers, governments, and the public to stay informed about the AI models deployed today. The disclosure should come with a self-certification that the lab is complying with the terms of their published Secure Development Framework.
- **Publish a System Card:** System cards or other documentation should summarize the testing and evaluation procedures, results and mitigations required (subject to appropriate redaction for information that could compromise public safety or the safety and security of the model). The system card should also be publicly disclosed at deployment, and updated if the model is substantially revised.
- **Protect Whistleblowers by Prohibiting False Statements:** Explicitly make it a violation of law for a lab to lie about its compliance with its framework. This clarification creates a clear legal violation that enables existing whistleblower protections to apply and ensures that enforcement resources are squarely focused on labs that have engaged in purposeful misconduct.
- **Transparency Standards:** A workable AI transparency framework should have a minimum set of standards so that it can enhance security and public safety while accommodating the evolving nature of AI development. Given that AI safety and security practices remain in their early stages, with frontier developers like Anthropic actively researching best practices, any framework must be designed for evolution. Standards should begin as flexible, lightweight requirements that can adapt as consensus best practices emerge among industry, government, and other stakeholders.

This transparency approach sheds light on industry best practices for safety and can help set a baseline for how responsible labs train their models, ensuring developers meet basic accountability standards while enabling the public and policymakers to distinguish between responsible and irresponsible practices. For example, the Secure Development Framework we describe here is akin to Anthropic's own Responsible Scaling Policy and others from leading labs (Google DeepMind, OpenAI, Microsoft), all of whom have already implemented similar approaches while releasing frontier models. Putting a Secure Development Framework transparency requirement into law would not

only standardize industry best practices without setting them in stone, it would also ensure that the disclosures (which are now voluntary) could not be withdrawn in the future as models become more powerful.

Views differ on whether and when AI models could pose catastrophic risks. Transparency requirements for Secure Development Frameworks and system cards could help give policymakers the evidence they need to determine if further regulation is warranted, as well as provide the public with important information about this powerful new technology.

As models advance, we have an unprecedented opportunity to accelerate scientific discovery, healthcare, and economic growth. Without safe and responsible development, a single catastrophic failure could halt progress for decades. **Our proposed transparency framework offers a practical first step:** public visibility into safety practices while preserving private sector agility to deliver AI's transformative potential.

News

Claude is now generally available in Xcode

Sep 16, 2025

News

Strengthening our safeguards through collaboration with US CAISI and UK AISI

Sep 13, 2025

News

Bringing memory to teams at work

Sep 12, 2025

Overview

Developer docs

Pricing

Amazon Bedrock

Google Cloud's Vertex AI

Console login

Learn

Courses

Connectors

Customer stories

Engineering at Anthropic

Events

Powered by Claude

Service partners

Startups program

Company

Anthropic

Careers

Economic Futures

Research

News

Responsible Scaling Policy

Security and compliance

Transparency

Help and security

Availability

Status

Support center

Terms and policies

Privacy choices

Privacy policy

Responsible disclosure policy

Terms of service: Commercial

Terms of service: Consumer

Usage policy