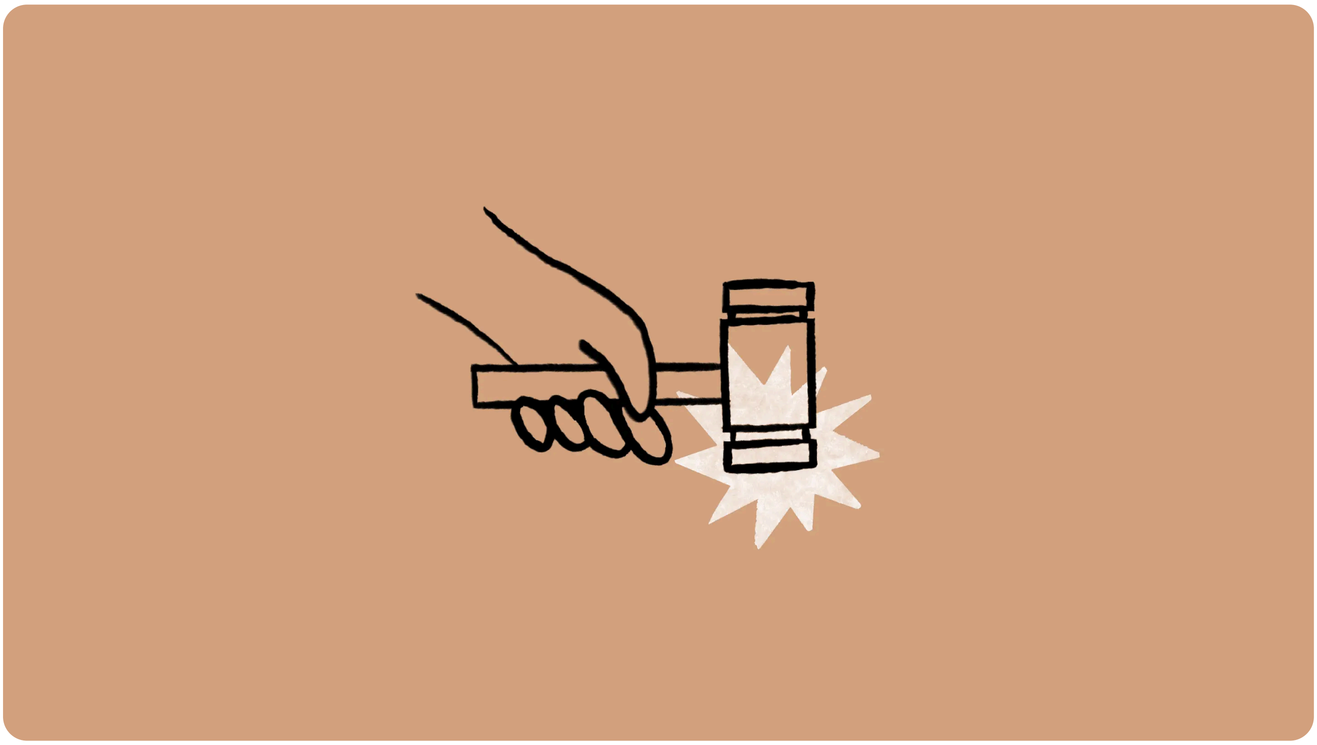
Announcements

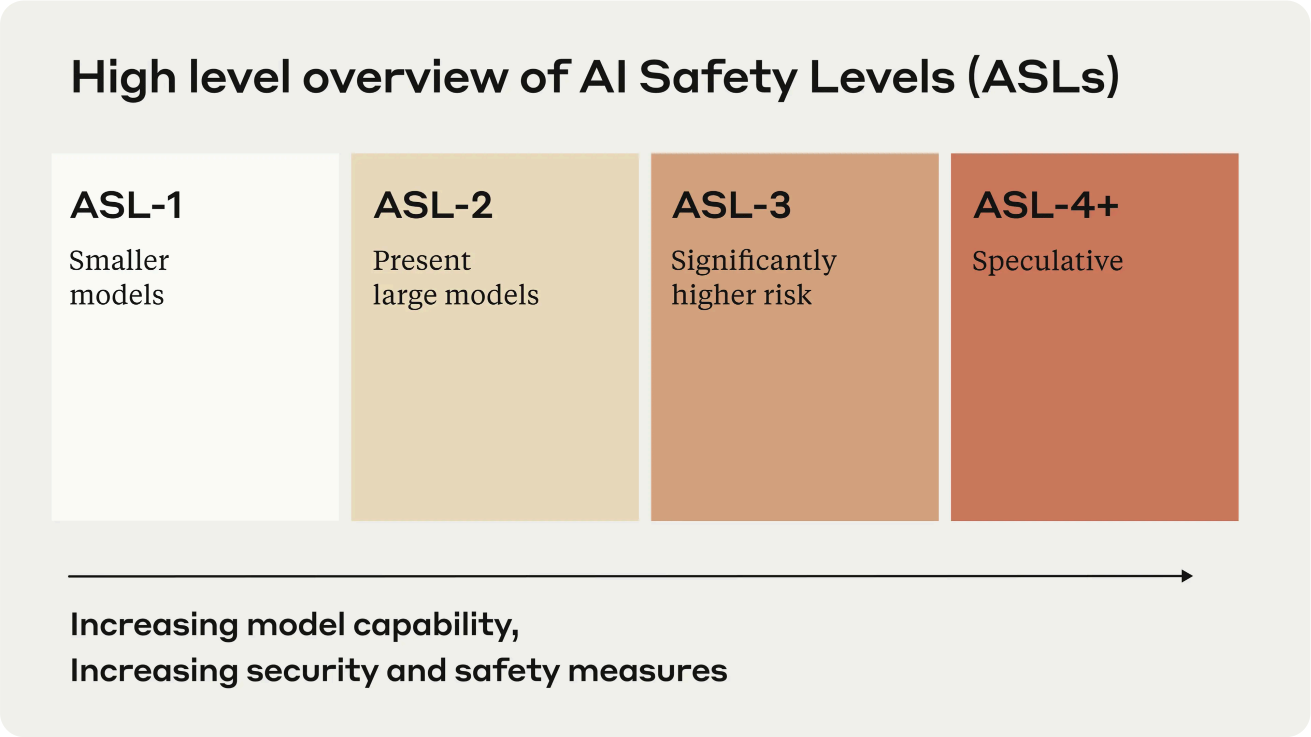
Anthropic's Responsible Scaling Policy

Sep 19, 2023 • 4 min read



Today, we're publishing our Responsible Scaling Policy (RSP) – a series of technical and organizational protocols that we're adopting to help us manage the risks of developing increasingly capable AI systems.

As AI models become more capable, we believe that they will create major economic and social value, but will also present increasingly severe risks. Our RSP focuses on catastrophic risks – those where an AI model directly causes large scale devastation. Such risks can come from deliberate misuse of models (for example use by terrorists or state actors to create bioweapons) or from models that cause destruction by acting autonomously in ways contrary to the intent of their designers.



High level overview of AI Safety Levels (ASLs)

ASL-1
Smaller models

ASL-2
Present large models

ASL-3
Significantly higher risk

ASL-4+
Speculative

Increasing model capability,
Increasing security and safety measures

Our RSP defines a framework called AI Safety Levels (ASL) for addressing catastrophic risks, modeled loosely after the US government's biosafety level (BSL) standards for handling of dangerous biological materials. The basic idea is to require safety, security, and operational standards appropriate to a model's potential for catastrophic risk, with higher ASL levels requiring increasingly strict demonstrations of safety.

A very abbreviated summary of the ASL system is as follows:

- ASL-1 refers to systems which pose no meaningful catastrophic risk, for example a 2018 LLM or an AI system that only plays chess.
- ASL-2 refers to systems that show early signs of dangerous capabilities – for example ability to give instructions on how to build bioweapons – but where the information is not yet useful due to insufficient reliability or not providing information that e.g. a search engine couldn't. Current LLMs, including Claude, appear to be ASL-2.
- ASL-3 refers to systems that substantially increase the risk of catastrophic misuse compared to non-AI baselines (e.g. search engines or textbooks) OR that show low-level autonomous capabilities.
- ASL-4 and higher (ASL-5+) is not yet defined as it is too far from present systems, but will likely involve qualitative escalations in catastrophic misuse potential and autonomy.

The definition, criteria, and safety measures for each ASL level are described in detail in the main document, but at a high level, ASL-2 measures represent our current safety and security standards and overlap significantly with our recent White House commitments. ASL-3 measures include stricter standards that will require intense research and engineering effort to comply with in time, such as unusually strong security requirements and a commitment not to deploy ASL-3 models if they show any meaningful catastrophic misuse risk under adversarial testing by world-class red-teamers (this is in contrast to merely a commitment to perform red-teaming). Our ASL-4 measures aren't yet written (our commitment is to write them before we reach ASL-3), but may require methods of assurance that are unsolved research problems today, such as using interpretability methods to demonstrate mechanistically that a model is unlikely to engage in certain catastrophic behaviors.

We have designed the ASL system to strike a balance between effectively targeting catastrophic risk and incentivising beneficial applications and safety progress. On the one hand, the ASL system implicitly requires us to temporarily pause training of more powerful models if our AI scaling outstrips our ability to comply with the necessary safety procedures. But it does so in a way that directly incentivizes us to solve the necessary safety issues as a way to unlock further scaling, and allows us to use the most powerful models from the previous ASL level as a tool for developing safety features for the next level.[1] If adopted as a standard across frontier labs, we hope this might create a "race to the top" dynamic where competitive incentives are directly channeled into solving safety problems.

From a business perspective, we want to be clear that our RSP will not alter current uses of Claude or disrupt availability of our products. Rather, it should be seen as analogous to the pre-market testing and safety feature design conducted in the automotive or aviation industry, where the goal is to rigorously demonstrate the safety of a product before it is released onto the market, which ultimately benefits customers.

Anthropic's RSP has been formally approved by its board and changes must be approved by the board following consultations with the Long Term Benefit Trust. In the full document we describe a number of procedural safeguards to ensure the integrity of the evaluation process.

However, we want to emphasize that these commitments are our current best guess, and an early iteration that we will build on. The fast pace and many uncertainties of AI as a field imply that, unlike the relatively stable BSL system, rapid iteration and course correction will almost certainly be necessary.

The full document can be read here. We hope that it provides useful inspiration to policymakers, third party nonprofit organizations, and other companies facing similar deployment decisions.

We thank ARC Evals for their key insights and expertise supporting the development of our RSP commitments, particularly regarding evaluations for autonomous capabilities. We found their expertise in AI risk assessment to be instrumental as we designed our evaluation procedures. We also recognize ARC Evals' leadership in originating and spearheading the development of their broader ARC Responsible Scaling Policy framework, which inspired our approach.

Footnotes

[1] As a general matter, Anthropic has consistently found that working with frontier AI models is an essential ingredient in developing new methods to mitigate the risk of AI.

News

Seoul becomes Anthropic’s third office in Asia-Pacific as we continue our international growth

Oct 23, 2025

News

Expanding our use of Google Cloud TPUs and Services

Oct 23, 2025

News

A statement from Dario Amodei on Anthropic's commitment to American AI leadership

Oct 21, 2025

Products

Claude

Claude Code

Max plan

Team plan

Enterprise plan

Download app

Pricing

Log in to Claude

Models

Opus

Sonnet

Haiku

Solutions

AI agents

Code modernization

Coding

Customer support

Education

Financial services

Government

Life sciences

Claude Developer Platform

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