Anthropic Responsible Scaling Policy (RSP) v2

ANTHROPIC-RSP-2024 · US · voluntary code

Summary

First-mover industry safety framework. Introduces the AI Safety Level (ASL) capability-tier vocabulary subsequently adapted by OpenAI Preparedness + DeepMind FSF. v2 (Oct 2024) refines ASL-3/ASL-4 capability thresholds, mandates pre-deployment capability evaluations, and commits to a Frontier Red Team. Seoul Frontier AI Safety Commitments signatory; cited by name in EU AI Office GPAI Code of Practice drafts. NOTE (iter-314): the RSP is a versioned-evolving artefact; this row pins v2 (Oct 2024) as the load-bearing reference, but Anthropic publishes incremental updates on the policy page. Citers tracking specific ASL-4 threshold language should confirm against the current version on anthropic.com — the catalog re-pins on the next Coverage Games event. Currency (2026-06-21): superseded as a reference by RSP v3.x (current v3.3, 2026-05-26) — v3.0 (24 Feb 2026) was a comprehensive rewrite that replaced the binding ASL hard-limit with a Frontier Safety Roadmap of publicly-declared targets plus Risk Reports and independent external review, so the v2 (Oct 2024) ASL-threshold language this row pins is now two major versions out of date.

At a glance

Adopted

2024-10-15

Effective

2024-10-15

Status

in force

Primary source

Anthropic Responsible Scaling Policy v2 (Oct 2024)

How to cite this article

APA

Policy Window. (2024). Anthropic Responsible Scaling Policy (RSP) v2 [Wiki article — Instrument]. https://policywindow.org/wiki/anthropic-rsp

CHICAGO

Policy Window. 2024. "Anthropic Responsible Scaling Policy (RSP) v2." Wiki article (Instrument). https://policywindow.org/wiki/anthropic-rsp.

HARVARD

Policy Window (2024) 'Anthropic Responsible Scaling Policy (RSP) v2', Wiki article — Instrument, available at: https://policywindow.org/wiki/anthropic-rsp.

OSCOLA

Policy Window, 'Anthropic Responsible Scaling Policy (RSP) v2' (Wiki article — Instrument, 2024) <https://policywindow.org/wiki/anthropic-rsp> accessed [date].

BIBTEX

```
@misc{policywindow-anthropic-rsp,
title  = {Anthropic Responsible Scaling Policy (RSP) v2},
author = {Policy Window},
year   = {2024},
howpublished = {Anthropic Responsible Scaling Policy v2 (Oct 2024)},
url    = {https://policywindow.org/wiki/anthropic-rsp},
note   = {Primary source: https://www.anthropic.com/news/announcing-our-updated-responsible-scaling-policy}
}
```