

# AI Safety Level 3 (ASL-3)

asl-3 · safety · concept

Source: <https://policywindow.org/wiki/asl-3>

Generated 2026-05-30T22:10:07 UTC

## Summary

A capability-based risk tier in Anthropic's Responsible Scaling Policy denoting models with the potential to substantially uplift CBRN attack capabilities or autonomous AI replication.

## At a glance

Used by

**3 instrument(s)**

Primary source

**Anthropic Responsible Scaling Policy v1.x**

Related concepts

**frontier-tier, systemic-risk, compute-threshold**

## Details

ASL-3 was introduced in Anthropic's Responsible Scaling Policy (RSP) framework. Triggering ASL-3 capability requires the model to demonstrate substantial uplift in chemical, biological, radiological, or nuclear (CBRN) weapons design beyond baseline internet resources, OR show signs of autonomous self-replication. ASL-3 status mandates specific deployment safeguards including red-team evaluations, restricted API access, and incident-response protocols. Comparable tiers exist in OpenAI's Preparedness Framework (high) and DeepMind's Frontier Safety Framework (Critical Capability Levels).

## How to cite this article

APA

Policy Window. (n.d.). AI Safety Level 3 (ASL-3) [Wiki article — Concept]. <https://policywindow.org/wiki/asl-3>

CHICAGO

Policy Window. n.d.. "AI Safety Level 3 (ASL-3)." Wiki article (Concept). <https://policywindow.org/wiki/asl-3>.

HARVARD

Policy Window (n.d.) 'AI Safety Level 3 (ASL-3)', Wiki article — Concept, available at: <https://policywindow.org/wiki/asl-3>.

OSCOLA

Policy Window, 'AI Safety Level 3 (ASL-3)' (Wiki article — Concept, n.d.) <<https://policywindow.org/wiki/asl-3>> accessed [date].

BIBTEX

```
@misc{policywindow-asl-3,  
  title = {AI Safety Level 3 (ASL-3)},  
  author = {Policy Window},  
  year = {n.d.},  
  howpublished = {asl-3 - safety},
```

```
url      = {https://policywindow.org/wiki/asl-3},
note    = {Primary source: https://www.anthropic.com/news/anthropics-responsible-scaling-policy}
}
```