Google DeepMind Frontier Safety Framework

DEEPMIND-FSF-2024 · US · voluntary code

Source: https://policywindow.org/wiki/deepmind-fsf

Generated 2026-11-19T01:05:54 UTC

Summary

Critical Capability Levels (CCL) regime spanning autonomy, biosecurity, cybersecurity, and persuasion domains. Distinct vocabulary from Anthropic ASL + OpenAI Preparedness — designed for cross-domain elicitation; each CCL triggers domain-specific mitigations including model-weight access controls + enhanced red-teaming. Seoul Frontier AI Safety Commitments signatory. Alphabet-published; effective across Google DeepMind frontier-model releases. NOTE (iter-314): the FSF is a versioned-evolving artefact; this row pins v1 (May 2024) as the load-bearing reference, but DeepMind publishes incremental updates on the deepmind.google blog. Citers tracking specific CCL definitions or mitigation requirements should confirm against the current published version — the catalog re-pins on the next Coverage Games event. Currency (2026-06-21): The catalog pins FSF v1 (May 2024), but DeepMind has since published v2.0 (4 Feb 2025), v3.0 (22 Sept 2025, adding a harmful-manipulation Critical Capability Level plus expanded misalignment and ML-R&D protocols), and v3.1 (17 Apr 2026, introducing Tracked Capability Levels); citers should confirm CCL definitions against the current version at deepmind.google/blog/strengthening-our-frontier-safety-framework/.

At a glance

Adopted

2024-05-17

Effective

2024-05-17

Status

in force

Primary source

Google DeepMind Frontier Safety Framework (May 2024)

How to cite this article

APA

Policy Window. (2024). Google DeepMind Frontier Safety Framework [Wiki article — Instrument]. https://policywindow.org/wiki/deepmind-fsf

CHICAGO

Policy Window. 2024. "Google DeepMind Frontier Safety Framework." Wiki article (Instrument). https://policywindow.org/wiki/deepmind-fsf.

HARVARD

Policy Window (2024) 'Google DeepMind Frontier Safety Framework', Wiki article — Instrument, available at: https://policywindow.org/wiki/deepmind-fsf.

OSCOLA

Policy Window, 'Google DeepMind Frontier Safety Framework' (Wiki article — Instrument, 2024) <https://policywindow.org/wiki/deepmind-fsf> accessed [date].

BIBTEX

```
@misc{policywindow-deepmind-fsf,
title  = {Google DeepMind Frontier Safety Framework},
author = {Policy Window},
year   = {2024},
howpublished = {Google DeepMind Frontier Safety Framework (May 2024)},
url    = {https://policywindow.org/wiki/deepmind-fsf},
note   = {Primary source: https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/}
}
```