Scalable Oversight

scalable-oversight · safety · concept

Source: https://policywindow.org/wiki/scalable-oversight

Generated 2026-11-19T01:06:05 UTC

Summary

The set of techniques for supervising AI systems whose outputs are too complex, too numerous, or too domain-distant for unaided human evaluators to judge correctness.

At a glance

Used by

4 instrument(s)

Related concepts

alignment, deceptive-alignment, capability-elicitation, red-team-evaluation

Primary source

Christiano, P., Shlegeris, B., Amodei, D. (2018), 'Supervising Strong Learners by Amplifying Weak Experts.'

Details

Scalable oversight addresses the 'who watches the watchers' problem at AI scale. When a model produces 10v outputs per day, or operates in a domain where the supervising human is not expert (e.g., novel mathematics, advanced biology), traditional human-in-the-loop review fails. Christiano et al. (2018) 'Supervising Strong Learners by Amplifying Weak Experts' is the foundational articulation. The agenda spans: (a) debate (two AIs argue, a human judges short transcripts — Irving et al. 2018); (b) iterated amplification (humans + assistants supervise stronger models, recursively — Christiano et al. 2018); (c) constitutional AI / RLAIF (rule-based or AI-feedback supervision in place of unscaled human labels — Bai et al. 2022, Anthropic); (d) weak-to-strong generalisation (Burns et al. 2023, OpenAI) — can a weak supervisor train a stronger model to behave well on tasks the weak supervisor cannot grade?

Governance relevance is direct. EU AI Act Art. 14 mandates 'human oversight' for high-risk systems; the article is written assuming bandwidth-feasible human review, which scalable-oversight literature argues breaks at frontier-model scale. UK AISI red-team commitments explicitly invoke scalable-oversight techniques. NIST AI RMF Govern 1.3 calls for documented oversight mechanisms but does not specify scalability requirements. The gap between regulatory 'human oversight' language and the technical reality of supervising super-human-domain outputs is one of the field's most-discussed governance-implementation gaps.

How to cite this article

APA

Policy Window. (n.d.). Scalable Oversight [Wiki article — Concept]. https://policywindow.org/wiki/scalable-oversight

CHICAGO

Policy Window. n.d.. "Scalable Oversight." Wiki article (Concept). https://policywindow.org/wiki/scalable-oversight.

HARVARD

Policy Window (n.d.) 'Scalable Oversight', Wiki article — Concept, available at: https://policywindow.org/wiki/scalable-oversight.

OSCOLA

Policy Window, 'Scalable Oversight' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/scalable-oversight> accessed [date].

BIBTEX

```
@misc{policywindow-scalable-oversight,
title  = {Scalable Oversight},
author = {Policy Window},
year   = {n.d.},
howpublished = {scalable-oversight — safety},
url    = {https://policywindow.org/wiki/scalable-oversight},
note   = {Primary source: https://arxiv.org/abs/1810.08575}
}
```