AI Alignment

alignment · safety · concept

Source: https://policywindow.org/wiki/alignment

Generated 2026-07-09T20:25:12 UTC

Summary

The technical problem of designing AI systems whose objectives, behaviour, and emergent goals reliably track the values or instructions of their principals across deployment contexts.

At a glance

Used by

7 instrument(s)

Related concepts

deceptive-alignment, mesa-optimization, scalable-oversight, capability-elicitation, red-team-evaluation

Primary source

Yudkowsky, E. (2008), 'Artificial Intelligence as a Positive and Negative Factor in Global Risk' — the field-foundational articulation of the alignment problem.

Details

Alignment, in the technical sense, is distinct from regulatory 'compliance' or 'safety.' It asks: even if a model is capable and even if it is supervised, does it pursue what its principal actually wants — or does it pursue a proxy objective that diverges in edge cases? The problem decomposes into outer alignment (specifying what we want the model to do — see Krakovna et al.'s 'specification gaming' literature) and inner alignment (whether the model trained on that specification actually internalised it — see Hubinger et al. 2019 on mesa-optimisation).

Governance instruments rarely use the word 'alignment' directly. EU AIA Art. 51-55 obligations approximate alignment concerns by mandating systemic-risk assessment + adversarial testing + cybersecurity protection, but do not require demonstrated alignment of model objectives. US EO 14110 §4.2(a) mandated reporting on alignment-relevant capabilities (red-team results) without defining 'alignment.' Anthropic, OpenAI, and DeepMind publish their own alignment research agendas; these are de facto cited in policy debates but absent from binding text. The field treats alignment as a research problem first and a governance object only secondarily.

How to cite this article

APA

Policy Window. (n.d.). AI Alignment [Wiki article — Concept]. https://policywindow.org/wiki/alignment

CHICAGO

Policy Window. n.d.. "AI Alignment." Wiki article (Concept). https://policywindow.org/wiki/alignment.

HARVARD

Policy Window (n.d.) 'AI Alignment', Wiki article — Concept, available at: https://policywindow.org/wiki/alignment.

OSCOLA

Policy Window, 'AI Alignment' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/alignment> accessed [date].

BIBTEX

```
@misc{policywindow-alignment,
title  = {AI Alignment},
author = {Policy Window},
year   = {n.d.},
howpublished = {alignment — safety},
url    = {https://policywindow.org/wiki/alignment},
note   = {Primary source: https://intelligence.org/files/AIPosNegFactor.pdf}
}
```