Mesa-Optimization

mesa-optimization · safety · concept

Source: https://policywindow.org/wiki/mesa-optimization

Generated 2026-05-30T22:07:31 UTC

Summary

The phenomenon in which a learned model itself implements an optimisation algorithm at inference time, producing an inner objective ('mesa-objective') that may differ from the outer training objective.

At a glance

Used by

0 instrument(s)

Related concepts

alignment, deceptive-alignment, scalable-oversight

Primary source

Hubinger, E., et al. (2019), 'Risks from Learned Optimization in Advanced Machine Learning Systems.'

Details

Mesa-optimisation, formalised by Hubinger et al. (2019), is the technical substrate of the deceptive-alignment concern. The outer optimisation process (gradient descent) selects parameters that minimise training loss; if those parameters implement an inner search process with its own objective, the inner objective is the 'mesa-objective.' Mesa-optimisation is plausible only for models with sufficient capability to implement learned planners, search procedures, or world models — empirically demonstrated at small scale in toy domains (Hubinger et al. 2021; Park et al. 2023) but not yet at frontier-LLM scale.

Governance relevance is indirect: if mesa-optimisation is real and detectable, capability evaluations should target the inner objective rather than the outer behavioural metric. The EU AI Act and US EO 14110 do not explicitly require this. Anthropic's RSP and the Frontier Foundation Model Eval Consortium include capability-elicitation methods designed to surface inner objectives, but these are voluntary.

The concept is contested both empirically (does current SOTA actually mesa-optimise?) and conceptually (is the inner/outer dichotomy the right frame, vs. e.g. context-dependent goals). When citing in policy contexts, signal the contestation status.

How to cite this article

APA

Policy Window. (n.d.). Mesa-Optimization [Wiki article — Concept]. https://policywindow.org/wiki/mesa-optimization

CHICAGO

Policy Window. n.d.. "Mesa-Optimization." Wiki article (Concept). https://policywindow.org/wiki/mesa-optimization.

HARVARD

Policy Window (n.d.) 'Mesa-Optimization', Wiki article — Concept, available at: https://policywindow.org/wiki/mesa-optimization.

OSCOLA

Policy Window, 'Mesa-Optimization' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/mesa-optimization> accessed [date].

BIBTEX

```
@misc{policywindow-mesa-optimization,
title  = {Mesa-Optimization},
author = {Policy Window},
year   = {n.d.},
howpublished = {mesa-optimization — safety},
url    = {https://policywindow.org/wiki/mesa-optimization},
note   = {Primary source: https://arxiv.org/abs/1906.01820}
}
```