The OTHER AI Alignment Problem: Mesa-Optimizers and Inner Alignment

Inner alignment is the problem of making sure that the goal an AI ends up pursuing is the same as the goal we optimized it for.

Machine learning uses an optimization algorithm called Stochastic Gradient Descent (SGD) to find algorithms which perform well according to some objective function. SGD is called the "base optimizer" and it finds learned algorithms that perform well according to the "base objective." A "mesa-optimizer" is a learned algorithm that is itself an optimizer. A "mesa-objective" is the objective of a mesa-optimizer. So what we call the inner alignment problem is making sure that if the AI is a mesa-optimizer, its mesa-objective is equal to the base objective.

As an analogy: natural selection can be seen as an optimization algorithm that 'designed' humans to achieve the goal of high genetic fitness, or, roughly, "have lots of descendants." However, humans no longer primarily pursue reproductive success; they instead use birth control while still attaining the pleasure that natural selection 'meant' as a reward for attempts at reproduction. This is a failure of inner alignment.

The inner alignment problem can be split into sub-problems like deceptive alignment, distribution shifts, and gradient hacking.

Related

- What is outer alignment?
- What is deceptive alignment?
- What is the difference between inner and outer alignment?
- What is goal misgeneralization?
- What are mesa-optimizers?
- What is the likelihood of inner misalignment?

Scratchpad

Removed

Inner alignment asks, "Is the model *trying to do* what humans have specified it should do?", or in other words, can we robustly aim our AIs at anything at all?

More specifically, inner alignment is the problem of ensuring that any mesa-optimizer (i.e. a trained machine learning system which is itself an optimizer) is aligned with the objective function of the training process.

You can have both inner and outer alignment failures together. It is not a dichotomy and often even experienced alignment researchers are unable to tell them apart. Ideally, we don't think of a dichotomy of inner and outer alignment that can be tackled individually but of a more holistic alignment picture that includes the interplay between both inner and outer alignment approaches.

The term was first defined in Risk from Learned Optimization:

> *We refer to this problem of aligning mesa-optimizers with the base objective as the inner alignment problem. This is distinct from the outer alignment problem, which is the traditional problem of ensuring that the base objective captures the intended goal of the programmers.*

This answer goes into more depth regarding the differences between inner and outer alignment.

Sources

- Defining capability and alignment in gradient descent
- "Inner Alignment Failures" Which Are Actually Outer Alignment Failures
- Inner alignment: what are we pointing at?
- Inner Alignment: Explain like I'm 12 Edition
- Explaining inner alignment to myself
- Comparing Four Approaches to Inner Alignment

- Hubinger, Evan; et. al. (May 2019) Risks from Learned Optimization
- Mikulik, Vladimir (Aug 2019) 2-D robustness
- Harris, Edouard (Nov 2020) Defining capability and alignment in gradient descent
- Hubinger, Evan (Nov 2020) Clarifying inner alignment terminology
- Filan, Daniel (Feb 2021) AXRP Ep.4 - Risks from Learned Optimization with Evan Hubinger
- Langosco, Lauro; jbkjr (Jun 2021) Objective Robustness and Inner Alignment Terminology
- Demski, Abram (July 2021) Re-Define Intent Alignment?
- Demski, Abram (July 2021) refactoring alignment #2
- Arike, Rauno (May 2022) Clarifying the confusion around inner alignment
- Demski, Abram (Sep 2022) Builder/Breaker for Deconfusion
- davidad (Dec 2022) Reframing inner alignment