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

The paper Risks from Learned Optimization in Advanced Machine Learning Systems makes the distinction between inner and outer alignment: Outer alignment means making the optimization target of the *training process* ("outer optimization target", e.g., the *loss* in supervised learning) aligned with what we want. Inner alignment means making the optimization target of the *trained system* ("inner optimization target") aligned with the outer optimization target. A challenge here is that the inner optimization target does not have an explicit representation in current systems, and can differ very much from the outer optimization target (see for example Goal Misgeneralization in Deep Reinforcement Learning).

See also this post for an intuitive explanation of inner and outer alignment.

Alternative phrasings

- What types of misalignment are there?

Related

- What is outer alignment?
- What is inner alignment?
- What is "Goal misgeneralization"?
- What is "reward misspecification"?