Read Hubinger's research agenda for more information.

It seems likely that deceptive agents are the default, so a key problem in alignment is how we can avoid deceptive alignment at every point in the training process. This seems to rely on being able to consistently exert optimization pressure against deception, which probably necessitates interpretability tools.

Hubinger's plan is "acceptability verification": have some predicate that precludes deception, and then check your model for this predicate at every point in training.

One idea for this predicate is making sure that the agent is myopic, meaning that the AI only cares about the current timestep, so there is no incentive to deceive, because the benefits of deception happen only in the future. This is operationalized as "return the action that your model of HCH would return, if it received your inputs."

Related

- What is inner alignment?
- What is deceptive alignment?
- What is the difference between inner and outer alignment?
- What does MIRI think about technical alignment?

Scratchpad

Taken down on 2024–05-02 due to Kabir mentioning it was out of date