We don't have an answer for this question yet. Would you like to write one?

Training AI Without Writing A Reward Function, with Reward Modelling

Alternative phrasings

-

Related

- What is reward modeling?
- What is reward hacking?
- What is reward tampering?
- What is reward design?

Scratchpad

Draft v0

Reward modeling involves learning a model of the reward function from human feedback. The aim is to understand and capture the intended objectives of the human evaluator without having to manually specify the reward function.

Recursive reward modeling extends the idea of reward modeling to a hierarchical structure where multiple levels of agents are involved. It involves decomposing a complex task into smaller subtasks and training agents to model the reward function for those subtasks. The reward models learned by the subtask agents are then used as feedback to train higher-level agents, forming a "tree" of reward modeling agents. The top-level agent is trained to optimize the original, hard task by leveraging the reward models of the subtask agents.

In recursive reward modeling, the focus is on decomposing a complex task into simpler subtasks and using reward modeling at each level to train agents that can perform those subtasks. This hierarchical structure allows for more efficient training and credit assignment, as well as the exploration of novel solutions that may not be apparent to humans. Recursive reward modeling is a specific instantiation of the broader concept of reward modeling, where the modeling process is applied recursively in a hierarchical manner.

Meta-Notes for self:

What is the difference between recursive reward modeling and IDA or HCH?