Techniques for enhancing human feedback

By Ajeya Cotra

Training powerful models to maximize simple metrics (such as quarterly profits) could be risky. Sufficiently intelligent models could discover strategies for maximizing these metrics in perverse and unintended ways. For example, the easiest way to maximize profits may turn out to involve stealing money, manipulating whoever keeps records into reporting unattainably high profits, capturing regulators of the industry to be allowed to ship shoddy products or avoid taxes, etc. More generally, the most effective path to maximizing a simple metric may involve acquiring enough power to tamper directly with whatever instruments or sensors are used to evaluate the metric, effectively deceiving and disempowering humans to do so.

It seems significantly safer if powerful models could be trained using something like human feedback, where human evaluators inspect a model's action and rate how good that action is likely to be all-things-considered, and the model is trained to take actions that humans would rate highly. Human feedback could potentially disincentivize some obviously-perverse strategies like "blatantly stealing money to maximize profits," and incentivize practices which could help maintain or improve human control like "explaining why a proposed action will be beneficial."

However, human feedback isn't fully adequate for supervising powerful models, especially if they take actions that are too complex for humans to understand. For example, even if blatant forms of theft are disincentivized, a sufficiently intelligent model trained with human feedback may still e.g. participate in various abstruse and complicated financial contracts which effectively constitute theft. On the other hand, if human evaluators simply penalize any action they don't understand, the models they train would be significantly less valuable than they could have been and may be outcompeted by models trained with outcome metrics like profit.

We are interested in ways to enhance or improve upon human feedback, so that humans can provide adequate feedback even in domains where models are more capable or knowledgable than humans, without resorting to training on outcome metrics. For example, projects in this space could explore questions like:

- How could human evaluators that are unfamiliar with a certain subject (like computer science or economics) effectively give feedback that incentivizes a model to accurately explain things about the subject?
- How could human evaluators effectively provide feedback to an RL agent acting in a virtual environment that is partially occluded from the evaluators, or operates based on internal dynamics that the evaluators don't understand?
- How could human evaluators effectively give feedback that incentivizes a model to accurately translate between English and a foreign language that the evaluators don't understand?

- How would a model behave if it's trained using a combination of an outcome metric and a human feedback signal? What happens if the outcome metric (e.g. "getting the most money in a negotiation game") incentivizes doing something undesired (e.g. "lying") that the human evaluators are unable to detect consistently?

We are seeking proposals for projects aiming to develop and test strategies for improving on the performance that could be achieved with naive human feedback in settings like these. Potential strategies that could be explored include:

- Training helper models that help human raters make better-informed evaluations of the model they are evaluating, e.g. by highlighting flaws or contradictions in the other model's statements, or fetching relevant information from the internet.
- Breaking down the evaluation of a complex task into smaller sub-questions that human raters have an easier time answering with confidence, improving the quality and reliability of their overall evaluation.
- Training on a curriculum of human evaluations given after different amounts of thoughts and reflection (e.g. 1 minute vs 10 minutes vs 1 hour), and testing whether the model robustly generalizes to human evaluations after a longer period of time (e.g. 5 hours).
- Developing automated quality control checks for human raters' judgments which throw out judgments that have an elevated probability of being incorrect, or boost the weight of judgments that have an elevated probability of being correct.

We are especially interested in proposals that focus on fine-tuning GPT-3 or another large generative language model, proposals that aim to use non-expert human raters to effectively evaluate a task that involves some subject matter expertise (e.g. legal advice, medical advice, programming, physics questions, etc), and proposals that systematically study how different techniques for enhancing human feedback may change the way performance scales as a function of the amount of human labor put into providing feedback.