Eliciting Latent Knowledge (ELK) is an open problem in AI safety. It asks how we can access information that an AI model "knows", but which its output doesn't include or make clear by default.

The problem statement from the original report by the Alignment Research Center (ARC):

> "Suppose we train a model to predict what the future will look like according to cameras and other sensors. We then use planning algorithms to find a sequence of actions that lead to predicted futures that look good to us.
>
> But some action sequences could tamper with the cameras so they show happy humans regardless of what's really happening. More generally, some futures look great on camera but are actually catastrophically bad.
>
> In these cases, the prediction model "knows" facts (like "the camera was tampered with") that are not visible on camera but would change our evaluation of the predicted future if we learned them. **How can we train this model to report its latent knowledge of off-screen events?"**

The hope is that a successful method for ELK would make it easier to tell if an AI is safe, or in what ways it might be unsafe.

ARC's original report includes a description of ARC's first-pass research strategy for solving ELK using a form of red teaming. The report goes on to describe a number of approaches to ELK that ARC considered and difficulties with each of them.

A summary of both ARC's original ELK report and the results of the 2022 ELK proposal contest is available here.

Related

- How is the Alignment Research Center (ARC) trying to solve Eliciting Latent Knowl…
- What is the Alignment Research Center (ARC)'s research strategy?
- If we solve ELK, how could we use that to align an AI?
- What is interpretability and what approaches are there?

Scratchpad

Sometimes, an AI model can "know" something that it doesn't communicate explicitly. For example, we might train models to predict what actions would lead to the outcomes we want captured on a video feed, and do those things. However, one way to capture good outcomes on a video feed is to tamper with the feed itself. In this case, the AI "knows" that the camera was tampered with, but might not tell us.

Eliciting Latent Knowledge (ELK) is the problem of accessing information that an AI model "knows", but that it doesn't communicate explicitly by default.

The hope is that a successful method for ELK would make it easier to tell if an AI is safe, or in what ways it might be unsafe.

ARC's original report includes a description of ARC's first-pass research strategy for solving ELK using a form of red teaming. The report goes on to describe a number of approaches to ELK that ARC considered and difficulties with each of them.

A summary of both ARC's original ELK report and the results of the 2022 ELK proposal contest is available here.