Research directions in AI control

Paul Christiano Follow
Dec 5, 2015 · 2 min read

What research would best advance our understanding of AI control?

I've been thinking about this question a lot over the last few weeks. This post lays out my best guesses.

My current take on AI control

I want to focus on existing AI techniques, minimizing speculation about future developments. As a special case, I would like to use minimal assumptions about unsupervised learning, instead relying on supervised and reinforcement learning. My goal is to find scalable approaches to AI control that can be applied to existing AI systems.

For now, I think that act-based approaches look significantly more promising than goal-directed approaches. (Note that both categories are consistent with using value learning.) I think that many apparent problems are distinctive to goal-directed approaches and can be temporarily set aside. But a more direct motivation is that the goal-directed approach seems to require speculative future developments in AI, whereas we can take a stab at the act-based approach now (though obviously much more work is needed).

In light of those views, I find the following research directions most attractive:

Four promising directions

- Elaborating on apprenticeship learning.
  Imitating human behavior seems especially promising as a scalable approach to AI control, but there are many outstanding problems.

- Efficiently using human feedback.
  The limited availability of human feedback may be a serious bottleneck for realistic approaches to AI control.

- Explaining human judgments and disagreements.
  My preferred approach to AI control requires humans to

understand AIs' plans and beliefs. We don't know how to solve the analogous problem for humans.

- Designing feedback mechanisms for reinforcement learning. A grab bag of problems, united by a need for proxies of hard-to-optimize, implicit objectives.

I will probably be doing work in one or more of these directions soon. I am also interested in talking with anyone who is considering looking into these or similar questions.

I'd love to find considerations that would change my view—whether arguments against these projects, or more promising alternatives. But these are my current best guesses, and I consider them good enough that the right next step is to work on them.

(This research was supported as part of the Future of Life Institute FLI-RFP-AI1 program, grant #2015–143898.)