The easy goal inference problem is still hard

Goal inference and inverse reinforcement learning

Paul Christiano Follow

Apr 11, 2015 · 5 min read

One approach to the AI control problem goes like this:

- Observe what the user of the system says and does.
- Infer the user's preferences.
- Try to make the world better according to the user's preference, perhaps while working alongside the user and asking clarifying questions.

This approach has the major advantage that we can begin empirical work today—we can actually build systems which observe user behavior, try to figure out what the user wants, and then help with that. There are many applications that people care about already, and we can set to work on making rich toy models.

It seems great to develop these capabilities in parallel with other AI progress, and to address whatever difficulties actually arise, as they arise. That is, in each domain where AI can act effectively, we'd like to ensure that AI can also act effectively in the service of goals inferred from users (and that this inference is good enough to support foreseeable applications).

This approach gives us a nice, concrete model of each difficulty we are trying to address. It also provides a relatively clear indicator of whether our ability to control AI lags behind our ability to build it. And by being technically interesting and economically meaningful now, it can help actually integrate AI control with AI practice.

Overall I think that this is a particularly promising angle on the AI safety problem.

Modeling imperfection

That said, I think that this approach rests on an optimistic assumption: that it's possible to model a human as an imperfect rational agent, and to extract the real values which the human is imperfectly optimizing. Without this assumption, it seems like some additional ideas are necessary.

To isolate this challenge, we can consider a vast simplification of the goal inference problem:

The easy goal inference problem: Given no algorithmic limitations and access to the complete human policy—a lookup table of what a human would do after making any sequence of observations—find any reasonable representation of any reasonable approximation to what that human wants.

I think that this problem remains wide open, and that we've made very little headway on the general case. We can make the problem even easier, by considering a human in a simple toy universe making relatively simple decisions, but it still leaves us with a very tough problem.

It's not clear to me whether or exactly how progress in AI will make this problem easier. I can certainly see how enough progress in cognitive science might yield an answer, but it seems much more likely that it will instead tell us "Your question wasn't well defined." What do we do then?

I am especially interested in this problem because I think that "business as usual" progress in AI will probably lead to the ability to predict human behavior relatively well, and to emulate the performance of experts. So I really care about the residual—what do we need to know to address AI control, beyond what we need to know to build AI?

Narrow domains

We can solve the very easy goal inference problem in sufficiently narrow domains, where humans can behave approximately rationally and a simple error model is approximately right. So far this has been good enough.

But in the long run, humans make many decisions whose consequences aren't confined to a simple domain. This approach can can work for driving from point A to point B, but probably can't work for designing a city, running a company, or setting good policies.

There may be an approach which uses inverse reinforcement learning in simple domains as a building block in order to solve the whole AI control problem. Maybe it's not even a terribly complicated approach. But it's not a trivial problem, and I don't think it can be dismissed easily without some new ideas.

Modeling "mistakes" is fundamental

If we want to perform a task as well as an expert, inverse reinforcement learning is clearly a powerful approach.

But in in the long-term, many important applications require AIs to make decisions which are *better* than those of available human experts. This is part of the promise of AI, and it is the scenario in which AI control becomes most challenging.

In this context, we can't use the usual paradigm—"more accurate models are better." A perfectly accurate model will take us exactly to human mimicry and no farther.

The possible extra oomph of inverse reinforcement learning comes from an explicit model of the human's mistakes or bounded rationality. It's what specifies what the AI should do differently in order to be "smarter," what parts of the human's policy it should throw out. So it implicitly specifies which of the human behaviors the AI should keep. The error model isn't an afterthought—it's the main affair.

Modeling "mistakes" is hard

Existing error models for inverse reinforcement learning tend to be very simple, ranging from Gaussian noise in observations of the expert's behavior or sensor readings, to the assumption that the expert's choices are randomized with a bias towards better actions.

In fact humans are not rational agents with some noise on top. Our decisions are the product of a complicated mess of interacting process, optimized by evolution for the reproduction of our children's children. It's not clear there is any good answer to what a "perfect" human would do. If you were to find any principled answer to "what is the human brain optimizing?" the single most likely bet is probably something like "reproductive success." But this isn't the answer we are looking for.

I don't think that writing down a model of human imperfections, which describes how humans depart from the rational pursuit of

fixed goals, is likely to be any easier than writing down a complete model of human behavior.

We can't use normal AI techniques to learn this kind of model, either—what is it that makes a model good or bad? The standard view—"more accurate models are better"—is fine as long as your goal is just to emulate human performance. But this view doesn't provide guidance about how to separate the "good" part of human decisions from the "bad" part.

So what?

It's reasonable to take the attitude "Well, we'll deal with that problem when it comes up." But I think that there are a few things that we can do productively in advance.

- Inverse reinforcement learning / goal inference research motivated by applications to AI control should probably pay particular attention to the issue of modeling mistakes, and to the challenges that arise when trying to find a policy better than the one you are learning from.
- It's worth doing more theoretical research to understand this kind of difficulty and how to address it. This research can help identify other practical approaches to AI control, which can then be explored empirically.