Human arguments and AI control

Explanation and AI control

Paul Christiano Follow

Nov 22, 2015 · 6 min read

Consider the definition:

> *An action is good to the extent that I would judge it to be good, after hearing a detailed explanation assessing its advantages, disadvantages, and alternatives.*

Contrast with:

> *An action is good to the extent that I would judge it to be good, after an extensive and idealized process of reflection.*

The second definition seems like a conceptually useful ingredient for AI control. The first definition would be a lot more actionable though; if we could make it work, it seems more promising as an ingredient of concrete AI control proposals. In this context, the explanations would be crafted by AI systems trained to produce compelling explanations.

For this to work, we would need to reliably distinguish good from bad actions based only on their explanations—explanations which were chosen to be maximally compelling, rather than based on notions of "honesty" or a desire to be helpful.

To achieve this goal, these explanations would need to be quite rich in at least two senses:

- These explanations may depend on unfamiliar concepts and complex reasoning, which *themselves* require explanation.
- These explanations need to explore not only an option but its alternatives, not only the evidence for a conclusion but the evidence against. They may look more like "arguments" than an "explanations," even though there is no explicit adversarial dynamic.

To make things a bit easier, the explanations need not be produced strategically by a single actor. Instead, different parts of the explanation could be produced to satisfy different goals.

It would be nice to know whether this kind of explanation is feasible even in principle. Will sophisticated AI systems be able to explain their reasoning in a way that meets this bar? Or is there an inherent gap between the conclusions that a system can reach and the conclusions that it can convincingly justify to someone else?

How to study AI explanation?

This problem is more challenging and interesting for AI systems that are more sophisticated than their users. It seems hard for these systems to give a full description of their reasoning to the user, in the same way that it can be difficult for a brilliant human expert to explain their reasoning to a layperson.

This problem is more approachable and (again) more interesting for AI systems that are able to communicate with humans and have a theory of mind. Without these capabilities, it seems hard to give a full description of their reasoning to the user, in the same way that it can be difficult for a 4 year old to explain their reasoning to anyone.

Between these two difficulties, explanation is hard to study in the context of existing systems. (Though we can study related problems about understandability and transparency, which themselves seem important and which I hope to discuss soon.)

Fortunately, we have an excellent toy model of this problem: humans.

The problem

Consider a group of experts in domain X, and a user with a minimal understanding of that domain. We would like to design a system that lets the user discover some fact about X or make a decision related to X. For example, the experts may be finance professionals “helping” a consumer decide what financial products to buy. Or the experts may be physicists “helping” an undergraduate make an accurate prediction about a novel high-energy physics experiment.

We are interested in the case where:

1. The experts don’t care about the user or about what is really true,
2. The user doesn’t have much time to think about the problem,
3. The domain is very complex, requiring detailed background knowledge, unfamiliar abstractions, and challenging reasoning,

and

4. The user wants to make decisions that are nearly as good as the decisions that the experts would make if they cared.

The only lever at the user's disposal is control over the experts' incentives. We'll assume that the user has more or less complete control over incentives, in the same way that the designers of AI systems have control over their training process.

As a simple and classical example, the user could elicit a recommendation from one expert, ask a second expert to argue against that recommendation, and then reward whichever one of them is more persuasive.

The question is: can we design systems that let the user reliably accomplish their task?

Examining the analogy

Our goal is to build systems that allow the user to extract the "right" answer from the experts, even in the presence of very large gaps in knowledge, time, and ability. Success would suggest that, at least if our AI systems are similar to the human experts, we could use this kind of explanation process to define the "right" action, without significant performance loss.

That looks good as far as it goes, but there are some reasons to expect AI systems to be unlike the human experts.

- AI systems which are in some respects very sophisticated may lack many capabilities of human experts. Even once we know that humans can evaluate explanations sufficiently rigorously, there is a further problem of actually building AI systems which are able to explain themselves.

Building systems that can actually explain themselves is a major and complementary research question, which seems worth studying in the context of existing AI algorithms. The usefulness of explanation for AI control rests on **both** the human's ability to evaluate explanations and the AI's ability to explain itself; we are only trying to study the former.

- In some respects the gap between human users and AI systems may be much larger than the gap between the user and the experts. And certainly the thought process of AI systems is likely to be much more unfamiliar. Even if our techniques can close the latter gap, they may not be able to close the former.

The explanation problem for humans is not the very hardest version of the explanation problem. If we could solve this version, we'd have more work to do in order to generalize to ever more alien cognitive styles. That said, if we were able to bridge a large gap in abilities with minimal reduction in performance, it would suggest that we had developed scalable frameworks for explanation.

With respect to capability gaps, note that experts with a lot of time are quite a bit more capable than laypeople with little time. At the point when AI systems have such a big lead over their human users, it seems likely that human contributions to AI control will no longer be relevant. So I'm inclined to think that "alien minds" is a bigger issue than "big capability gaps."

- Human experts may "play nice," following usual norms of civil discourse even when violating them would technically better suit the incentives described by the user. More subtly, human experts may simply not be very good at deception. (We may also be able to build AI systems that aren't very good at deception, but this would require some extra work.)

This is something to watch out for and to try avoid; see the discussion in the next section.

Working on the problem

Setup

In order to test approaches to this problem, all we need are a few experts in a domain, a judge who doesn't understand that domain (or a sequence of such judges), and a challenging question for the judge to attempt to answer. These ingredients are readily available. We can evaluate performance by comparing the answers the judge comes up with to the answers the experts would have given, or in some cases by appeals to ground truth or to more authoritative authorities.

It's not clear what domains/experts/questions are most fruitful, and there are other degrees of freedom about how to set up the exercise; but the easiest way to sort these things out may be to try it (with a relatively short timescale, such as 2-8 hours) and see.

One non-trivial issue is implementing the expert's incentives. One could start with play money, treating the exercise as a game which the experts try in good faith to win. Of course, if the experts are paid rather than being volunteers, then there is a more straightforward way to determine their incentives. But realistically I think that the good faith efforts will be more important than incentives. We can get extra mileage by having the same experts play the game against varying users, increasing the probability that the experts identify any particular effective strategy, and by having retrospective collaborative efforts to improve the expert's strategies.

Approaches

The bigger question may be: are there promising approaches to this problem that are worth exploring?

Even though this seems like a question of natural interest, I don't really know of a literature on it (though it would be great to find one). It seems like trying random things and seeing how they would work would be pretty informative.

I've written about one possible mechanism, which I am very interested to test. That proposal also has many possible improvements, and in general I expect the whole thing to change a lot after making contact with reality.