Learn policies or goals?

Paul Christiano Follow
Apr 21, 2015 · 4 min read

I've recently proposed training agents to make the decision we would most approve of, rather than to rationally pursue the outcomes we most desire.

Many researchers object to this idea, based on the practical observation that learning goals seems easier and more powerful than directly learning good policies.

I don't think this objection actually undermines approval seeking; I think that goal inference and approval seeking are complements rather than substitutes.

I'll justify this claim in the context of approval-seeking and inverse reinforcement learning. But the examples are chosen for concreteness, and the basic point is more general.

Some objections

I've encountered three objections of this flavor:

1. The simplest representation of a policy is often in terms of a goal and a world model. So learning goals can be statistically and computationally easier than learning policies directly.
2. There is already plenty of data to form good models of the world and human goals, whereas training policies directly requires expensive feedback about individual decisions, or domain-specific expert demonstrations.
3. Approval-directed decisions are never better than the judgment of the human overseer, but we would ultimately like to make better decisions than any human.

My responses to these objections:

1. Maximizing approval is not (much) harder than inverse reinforcement

learning

For concreteness, imagine that **you** are in the position of the approval-directed agent, trying to find actions that will earn the approval of an alien overseer.

Here’s one simple strategy: observe the behavior of the aliens, use an inverse reinforcement learning algorithm to try to figure out what they value, and then take actions that help them get what they seem to want.

This is especially plausible if the aliens provided a good inverse reinforcement learning algorithm as a hint to help you get started. Then you could perform well even if you couldn’t discover such an algorithm on your own.

You might not start off using this strategy. And you might eventually find a better strategy. But over the long run, you’ll probably earn approval at least as well as if you pursued this simple strategy.

2. Maximizing approval doesn’t require (much) extra training data

It’s easier to get training data about what actions lead to what outcomes, or about human goals, then to get feedback on a large number of an AI’s decisions.

But the argument from the last section implies this isn’t a big deal: a semi-supervised learner can use observations about cause-and-effect or about the overseer’s values to help it make predictions about what actions will be approved of.

The approval-directed agent may still need some training data about approval, but not much: just enough to confirm a few hypotheses like the “the overseer approves of actions that lead to desired outcomes,” and “the overseer takes actions that lead to desired outcomes.” Then the definition of “desired outcomes” can be inferred as latent structure, mostly using observations of humans’ behavior across a wide range of contexts.

(The first of these hypotheses is actually a special case of the second. A human can infer both of them without seeing any training data about approval at all, and a good predictor could probably do the same.)

3. Maximizing approval leads to good outcomes, even with smart agents

Using approval-direction introduces a new problem: how can the human overseer determine what actions are good? This is especially problematic if the approval-directed agent is much smarter than the human overseer.

Explicit bootstrapping seems promising: rather than having a human overseer judge an action on their own, we allow them to interact with another approval-directed agent. As long as the (Human+AI) team is smart enough to evaluate the AI's actions, we have grounds for optimism.

(This kind of bootstrapping is central to my optimism about approval-directed agents.)

Inverse reinforcement learning offers an alternative approach to crossing the "human-level" boundary. It aims to model human decision-making as consisting of a rational part and a noisy/irrational/bounded part. We can then ignore the irrational part and use better algorithms and a better world model for the rational part. This is intuitively appealing, but I think that it remains to be seen whether it can work.

So why bother with the indirection?

I've argued that an approval-directed agent will do at least as well as an IRL agent, because it can always use IRL as a technique to earn approval. But then why not just start with IRL?

An approval-directed agent regards IRL as a means to an end, which can be improved upon as the agent learns more about humans. For example, an approval-directed agent recognizes that the learned utility function is only an approximation, and so it can deviate from that approximation if it predicts that the users would object to the consequences of following it.

At the meta level, when we design algorithms for IRL, we have some standard in mind that lets us say that one design is good and another is bad. By using approval-direction we codify that standard (an algorithm is good if it lets the AI find actions with high approval), so that the AI can do the same kind of work that we are doing when we design an IRL agent.

Of course it may also turn out that other techniques are needed to find good actions. If an approval-directed agent can find these techniques, then it can use them instead of or in addition to approaches based on goal inference.

I care because I think that a lot of work is still needed in order to scale IRL to very intelligent agents. I’m also not optimistic about finding the “right” answer any time soon (I don’t know if there is a “right” answer). Being able to say what we are trying to do with an IRL algorithm—and being able to leave parts of the problem to a clever AI rather than needing to solve the whole thing in advance—may improve our prospects significantly.