Act-based agents

Paul Christiano Follow
Nov 30, 2015 · 4 min read

I've recently discussed three kinds of learning systems:

- Approval-directed agents which take the action the user would most approve of.
- Imitation learners which take the action that the user would tell them to take.
- Narrow value learners which take the actions that the user would prefer they take.

These proposals all focus on the short-term instrumental preferences of their users. From the perspective of AI control I think this is the interesting aspect that deserves more attention.

Going forward I'll call this kind of approach "act-based" unless I hear something better (credit to Eliezer), and I'll call agents of this type "act-based agents."

Robustness

Act-based agents seem to be robust to certain kinds of errors. You need only the vaguest understanding of humans to guess that killing the user is: (1) not something they would approve of, (2) not something they would do, (3) not in line with their instrumental preferences.

So in order to get bad outcomes here you have to really mess up your model of what humans want (or more likely mess up the underlying framework in an important way). If we imagine a landscape of possible interpretations of human preferences, there is a "right" interpretation that we are shooting for. But if you start with a wrong answer that is anywhere in the neighborhood, you will do things like "ask the user what to do, and don't manipulate them." And these behaviors will eventually get you where you want to go.

That is to say, the "right" behavior is surrounded by a massive crater of "good enough" behaviors, and in the long-term they all converge to the same place. We just need to land in the crater.

Human enhancement

All of these approaches have a common fundamental drawback: they only have as much foresight as the user. In some sense this is why they are robust.

In order for these systems to behave wisely, the user has to actually *be* wise. Roughly, the users need to be intellectual peers of the AI systems they are using.

This may sound quite demanding. But after making a few observations, I think it may be a realistic goal:

- The user can draw upon every technology at their disposal—including other act-based agents. (This is discussed more precisely here under the heading of "efficacy.")
- The user doesn't need to be quite as smart as the AI systems they are using, they merely need to be within striking distance. For example, it seems fine if it takes a human a few days make a decision, or to understand and evaluate a decision, that an AI can make in a few seconds.
- The user can delegate this responsibility to other humans whom they are willing to trust (e.g. Google engineers), just like they do today.

In this story the capabilities of humans grow in parallel with the capabilities of AI systems, driven by close interaction between the two. AI systems do not pursue explicitly defined goals, but instead help the humans do whatever the humans want to do at any given time. The entire process remains necessarily comprehensible to humans—if humans can't understand how an action helps them achieve their goals, then that action doesn't get taken.

In speculations about the long-term future of AI, I think this may be the most common positive vision. But I don't think there has been much serious thinking about what this situation actually looks like, and certainly not much thinking about how to actually realize such a vision.

Note that the involvement of actual of humans is not intended as a *very* long-term solution. It's a solution built to last (at most) until all contemporary thinking about AI has been thoroughly obsoleted—until the capability of society is perhaps ten or a hundred times

greater than it is today. I don't think there is a strong case for thinking much further ahead than that.

What is "narrow" anyway?

There is clearly a difference between act-based agents and traditional rational agents. But it's not entirely clear what the key difference is.

Consider a machine choosing a move in a game of chess. I could articulate preferences over that move (castling looks best to me), over its consequences (I don't want to lose the bishop), over the outcome of the game (I want to win), over immediate consequences of that outcome (I want people to respect my research team), over distant consequences (I want to live a fulfilling life).

We could also go the other direction and get even narrower: rather than thinking about preferences over moves we can think about preferences over particular steps of the cognitive strategy that produces moves.

As I advance from "narrow" to "broad" preferences, many things are changing. It's not really clear what the important differences are, what exactly we mean by "narrow" preferences, at what scales outcomes are robust to errors, at what scales learning is feasible, and so on. I would like to understand the picture better.

The upshot

Thinking about act-based agents suggests a different (and in my view more optimistic) picture of AI control. There are a number of research problems that are common across act-based approaches, especially related to keeping humans up to speed, and I think that for the moment these are the most promising directions for work on AI control.