Ambitious vs. narrow value learning

Paul Christiano Follow
Oct 4, 2015 · 5 min read

Suppose I'm trying to build an AI system that "learns what I want" and helps me get it. I think that people sometimes use different interpretations of this goal. At two extremes of a spectrum of possible interpretations:

- The AI learns my preferences over (very) long-term outcomes. If I were to die tomorrow, it could continue pursuing my goals without me; if humanity were to disappear tomorrow, it could rebuild the kind of civilization we would want; *etc*. The AI might pursue radically different subgoals than I would on the scale of months and years, if it thinks that those subgoals better achieve what I really want.
- The AI learns the narrower subgoals and instrumental values I am pursuing. It learns that I am trying to schedule an appointment for Tuesday and that I want to avoid inconveniencing anyone, or that I am trying to fix a particular bug without introducing new problems, *etc*. It does not make any effort to pursue wildly different short-term goals than I would in order to better realize my long-term values, though it may help me correct some errors that I would be able to recognize as such.

I think that many researchers interested in AI safety per se mostly think about the former. I think that researchers with a more practical orientation mostly think about the latter.

The ambitious approach

The maximally ambitious approach has a natural theoretical appeal, but it also seems quite hard. It requires understanding human preferences in domains where humans are typically very uncertain, and where our answers to simple questions are often inconsistent, like how we should balance our own welfare with the welfare of others, or what kinds of activities we really want to pursue vs. enjoying in the moment. (It seems unlikely to me that there is a unified notion of "what I want" in many of these cases.) It also requires extrapolation to radically unfamiliar domains, where we will need to make decisions about issues like population ethics, what kinds of creatures do we care about, and unforeseen new technologies.

I have written about this problem, pointing out that it is unclear how you would solve it even with an unlimited amount of computing power. My impression is that most practitioners don't think of this problem even as a long-term research goal—it's a qualitatively different project without direct relevance to the kinds of problems they want to solve.

The narrow approach

The narrow approach looks relatively tractable and well-motivated by existing problems. We want to build machines that helps us do the things we want to do, and to that end they need to be able to understand what we are trying to do and what instrumental values guide our behavior. To the extent that our "preferences" are underdetermined or inconsistent, we are happy if our systems at least do as well as a human, and make the kinds of improvements that humans would reliably consider improvements.

But it's not clear that anything short of the maximally ambitious approach can solve the problem we ultimately care about. A sufficiently clever machine will be able to make long-term plans that are significantly better than human plans. In the long run, we will want to be able to use AI abilities to make these improved plans, and to generally perform tasks in ways that humans would never think of perform them—going far beyond correcting simple errors that can be easily recognized as such.

In defense of the narrow approach

I think that the narrow approach probably takes us much further than it at first appears. I've written about these arguments before, which are for the most part similar to the reasons that approval-directed agents or directly mimicking human behavior might work, but I'll quickly summarize them again:

Instrumental goals

Humans have many clear instrumental goals like "remaining in effective control of the AI systems I deploy," "acquiring resources and other influence in the world," or "better understanding the world and what I want.” A value learner may able to learn robust preferences like these and pursue those instrumental goals using all of its ingenuity. Such AI’s would not necessarily be at a significant disadvantage with respect to normal competition, yet the resources they acquired would remain under meaningful human control (if that’s what their users would prefer).

This requires learning robust formulations of concepts like “meaningful control,” but it does not require making inferences about cases where humans have conflicting intuitions, nor considering cases which are radically different from those encountered in training —AI systems can continue to gather training data and query their users even as the nature of human-AI interactions changes (if that’s what their users would prefer).

Process

Even if we can’t infer human preferences over very distant objects, we might be able to infer human preferences well enough to guide a process of deliberation (real or hypothetical). Using the inferred preferences of the human could help eliminate some of the errors that a human would traditionally make during deliberation. Presumably these errors run counter to a deliberator’s short-term objectives, if those objectives are properly understood, and this judgment doesn’t require a direct understanding of the deliberator’s big-picture values.

This kind of error-correction could be used as a complement to other kinds of idealization, like providing the human a lot of time, allowing them to consult a large community of advisors, or allowing them to use automated tools.

Such a process of error-corrected deliberation could itself be used to provide a more robust definition of values or a more forward looking criterion of action, such as “an outcome/action is valuable to the extent that I would/did judge it valuable after extensive deliberation.”

Bootstrapping

By interacting with AI assistants, humans can potentially form and execute very sophisticated plans; if so, simply helping them achieve their short-term goals may be all that is needed. For some discussion of this idea, see these three posts.

Conclusion

I think that researchers interested in scalable AI control have been too quick to dismiss “narrow” value learning as unrelated to their core challenge. Overall I expect that the availability of effective narrow value learning would significantly simplify the AI control problem even for superintelligent systems, though at the moment we don’t understand the relationship very well.

(Thanks to Andreas Stuhlmüller and Owain Evans for helpful discussion.)