The state of the steering problem

Paul Christiano Follow

Apr 10, 2015 · 5 min read

The steering problem asks: given some powerful AI capabilities, how can we engineer a system that is both efficient and aligned with human interests?

Here's what I think we know about the steering problem right now:

- **Given a question-answering system**, that can answer any precisely posed question as well as a human, I think we have a reasonable candidate solution. But the proposed solution is hard-to-test, relies on the sensible behavior of humans in very exotic situations, and is hard to adapt to more realistic capabilities.
- **Given a very powerful predictor**, I think we have a promising but unsatisfying solution. The big problem is that requires a *very* powerful predictor. The big advantage is that the solution can be tested thoroughly, and the predictor can be tested and trained directly on "important" predictions (rather than relying on transfer learning).
- **Given a reinforcement learner or supervised learner**, I think we have a superficially plausible solution. I have little idea whether it would really work. A solution to the steering problem under these assumptions would imply a solution under essentially any other reasonable assumptions.

Overall, this problem seem much easier than I anticipated. I feel more like "It's easy to see how things could go wrong," rather than "It's hard to see how things could go right."

I think these approaches are mostly common-sensical, and are in some sense an evasion of the more exotic issues that will need to be resolved eventually. ("What do we really want?", "what standards of reasoning should we ultimately accept?", and so on.) But in fact I think we have a good chance of evading these exotic issues for now, postponing a resolution until they no longer seem so exotic.

Open problems

Better solutions for reinforcement learning. My solution for reinforcement learning is definitely dubious. It would be great find new approaches, find new problems with the existing approach, better patch known problems with the existing approach, or try to find a more robust/reliable way to reason about possible solutions.

I'm confident that there is room for significant improvement, though I'm very unsure how good a solution we can ultimately find.

More conservative assumptions. All of these solutions make a number of "nice" assumptions. For example, I assume that the training error of our algorithms is either very small, or mostly incurred very early, or else that there is no small set of "make or break" decisions. But can we design algorithms that are robust to a modest number of adversarial failures at any point during their execution? (Note that adversarial failures can be correlated across different AIs, and that there are a number of reasons that such correlations might arise.) Or can we articulate plausible guarantees for our algorithms that rule out problematic behaviors?

Another mild assumption is that a modest amount of human labor is available to oversee AIs (we aren't trying to make an AI that can reconstruct civilization after an apocalypse). Removing this assumption is also an interesting problem—not so much because the scenario itself is particularly plausible, but because it could lead to much more robust solutions.

Move on. I think it may be helpful to specifically ask "Supposing that we can solve the steering problem, how can things still go wrong?" For example, we still need to avoid undesirable internal behavior by pieces of a system optimized for instrumental purposes. And we wouldn't be happy if our RL agent decided at a key moment that it really cared about self-actualization rather than getting a high reward. (It wouldn't be completely unheard of: humans were selected for reproduction, but often decide that we have better things to do.)

Are these serious problems? Are there other lurking dangers? I don't really know. These questions are more closely tied up with empirical issues and the particular techniques used to produce AI.

White box solutions. *(See next section.)* All of these solutions are "black box" approaches. It would be good to find white box solution in any model, under any assumptions. That is, to implement a white

box solution using *any* well-defined capability, or even infinite computing power.

To formalize the "white-box" requirement, we can try to implement the preferences of uncooperative agents, or work under other pessimistic assumptions that make black box approaches clearly unworkable.

Along similar lines, could we design a system that could efficiently create a good world even if its operators were unaging simulations of young children? Or a dog? Are these questions meaningful? If we know that a solution doesn't work or isn't meaningful for sufficiently immature or underdeveloped humans, can we really suppose that we are on the right side of a magical threshold?

Black box vs. white box methods

(This section's dichotomy is closely related to, but different from, Wei Dai's here.)

All of these solutions use human judgment as a "black box:" we define what the right behavior is by making reference only to what humans would do or say under appropriate conditions. For example, we think of someone's judgment as a "mistake" if they would change it after thinking about it enough and having it explained to them.

A different approach is to treat human behavior as a "white box:" to reason about *why* a human made a certain decision, and then to try figure out what the human really wanted based on an understanding. For example, we might say that someone's judgment is a "mistake" by looking at the causal process that produced the judgment, identifying some correspondence between that process and actual facts about the world, and noticing possible inconsistencies.

White box approaches seem more intuitively promising. Inverse reinforcement learning aims to model human behavior as a goal-directed process perturbed by noise and error, and to use the extracted goals to guide an AI's decisions. Eliezer describes an analogous proposal in *Creating Friendly AI*; I doubt he stands by the proposal, but I believe he does stand by his optimism about white box approaches.

Black box approaches suffer from some clear disadvantages. For example, an AI using one might "know" that we are making a mistake, yet still not care. We can try minimize the risk of error (as

we usually do in life), but it would be nice to do so in a more principled way. There are also some practical advantages: white box approaches extract motives which are *simpler* than the human they motivate, while black box approaches extract "motives" which may be much more complex.

That said, I don't yet see how to make a white box solution work, even in principle. Even given a perfectly accurate model of a person, and an unlimited amount of time to think, I don't know what kind of algorithm would be able to classify a particular utterance as an error. So for now I mostly consider this a big open question.