Modeling AI control with humans

Paul Christiano Follow

Jan 29, 2016 · 7 min read

I've been trying to build an aligned AI out of reward-maximizing modules. A successful scheme could also be applied to a simple toy problem involving greedy humans, which might be more intuitive.

This is similar to toy problems I've written about before (1, 2, 3, 4), but I think this is a cleaner and stronger statement.

The setup

Suppose that you have access to a human-simulation, and can run as many simulations as available computing hardware allows.

You can give the simulation input over a binary channel, and receive its output over a binary channel. At any time you can give the simulation a score in $[0, 1] \cup \{\perp\}$. At any time you can take a snapshot of the current simulation or stop the simulation, and you can restart simulations from any saved snapshot.

The simulated human will, to the best of their abilities, maximize:

$\mathbb{E}$[the next score I receive | I next receive a non-$\perp$ score].

The question is: using simulations of this kind, can we build an effective composite system aligned with our interests?

If we want to be concrete/colorful, we can give some detail to the simulated human. It's just a normal person (other than their unusual goal of score-maximization), living in a world much like ours. The communication channel is implemented by a special computer that the chosen human has access to. After each reward signal, the human is given some time to rest, recover, and enjoy themselves, and then we restart the state of the rest of the world.

The benchmark

To evaluate our performance, we compare to a hypothetical benchmark in which the simulated human is instead motivated to be maximally helpful to whoever is currently on the other end of the communication channel.

In the benchmark scenario, simulations cost the same amount and are identical in every way except for their motivations. We have the same ability to stop, start, and snapshot simulations.

Our goal is to build a system, out of our collection of reward-driven simulations, that is nearly as useful to us as this benchmark. Ideally the gap in usefulness would be a small additive constant. But I'd also be happy with a small multiplicative constant (e.g. it takes 10% longer to get the same tasks done), or even a moderate multiplicative constant (e.g. it takes 3x longer to get the same tasks done). And I'd be interested in (though not happy with) a large multiplicative constant (e.g. it takes 1000x longer to get the same tasks done).

Fitting AI control into this framework

Most of the schemes I've written about over the last year fit into this simple model, and I think that it's a useful intuition pump for understanding these schemes. For example:

- Approval-directed agents
- Meeting halfway
- Synthesizing training data
- Symbolic reasoning
- Active learning with opaque predictors

I think that a solution to this problem would be an extremely promising step for AI control. Conversely, as long as we can't solve this problem, I think that it is hard to argue that the AI control problem is totally resolved, since we could in fact find ourselves with capabilities similar to the greedy-human-simulation.

Examining the analogy

There are two natural questions:

- If we had a working approach to AI control, would it solve this toy problem?
- If we could solve this toy problem, would we have a working approach to AI control?

In the next section I'll explore a few key disanalogies that I think are worth having in mind, but which are easy to capture by modifying the setup. But first I want to compare existing AI techniques to this basic model.

- Supervised learning fits into this model very well. During training we give the human a sequence of examples, with scores determined by the loss of their classification. After training, we take a snapshot of the simulation's state and then use that snapshot indefinitely.
- Online learning also fits into the model; in this case we simply continue to do training instead of using a snapshot.
- Episodic reinforcement learning fits into the model very well. During an episode we provide the reward as an ordinary observation; at the end of the episode we compute the average discounted reward and provide it as the score.
- Generative models aren't analogous to a single human-simulation, because it is not clear how to compute the reward. But many successful techniques for generative modeling, such as variational autoencoders or adversarial networks, are made out of a small number of components which do fit in this framework.
- Similarly, unsupervised learning doesn't fit directly into this framework because it is not clear what objective is being maximized. But most approaches that maximize an objective would fit in this framework, since e.g. we can ask the human to make predictions about upcoming data or generate new data (as in the previous post).

Some other tasks don't fit well into this framework, e.g. if much of the work is spent on specifying models or producing domain-specific algorithms that encode both how to solve the task and what the intended behavior is. For example, many approaches to robotics don't fit into this framework very well.

Overall, I feel like most existing learning algorithms fit into this framework, and so we should expect a robust solution to the AI control problem to also apply to the toy problem.

One objection is that over time researchers are thinking of increasingly clever ways to put together individual reward-maximizing modules (e.g. variational autoencoders). It may be that solving the AI control problem requires using new approaches that haven't yet been invented.

I don’t think this is a strong objection. There may be clever tricks that we don’t yet know for using reward-maximizing components to do cool things. But that just makes the toy problem richer and more interesting, and makes it more likely that progress on the toy problem will be interesting to AI researchers more broadly.

The reverse question—if we solve the toy problem, will it apply to the real AI control problem?—is a lot murkier. I would consider it a really good first step, and I think that it would immediately lead to a reasonable candidate solution to the AI control problem, since we could simply replace each human simulation with a reinforcement learner. But we’d need to do more work to get the that solution to actually work in practice, and there is a lot of room for new problems that don’t appear in the simple idealization.

Disanalogies

There are some very important disanalogies between the toy problem and the real AI control problem. Fortunately, it is pretty easy to modify the toy problem to fix them.

AI may be subhuman in many respects

In a solution to the toy problem, we may be tempted to rely on the fact that our simulated human is able to do anything that a human can do.

But a realistic AI system may be much worse than humans at some tasks. This could be true even for AI systems that are powerful enough to radically transform the world.

Particularly troubling is that an AI system might be especially bad at understanding humans, or human values. Humans may be especially good at some kinds of reasoning-about-humans, since we can e.g. put ourselves in each other’s shoes and engage a bunch of dedicated cognitive machinery produced by evolution.

An AI system that was very good at computational tasks but very bad at humans might pose distinctive dangers. Similarly, other big mismatches could lead to other kinds of trouble.

An ideal solution to the toy problem would continue to work if the human simulation is significantly subhuman in some respects. It might rest on *some* minimal set of capabilities, but we would like the minimum to be as minimal as possible. We should be especially wary of abilities like our better understanding of humans that seem plausibly human-specific.

AI may be superhuman in many respects

On the flip side, our AI systems may eventually be radically better than humans at many tasks. Ideally a toy problem would continue to work for *very* powerful human-simulations. To illustrate the point colorfully, rather than imagining a single human interacting with us, we could imagine an entire civilization of geniuses thinking for years during each second of simulation.

An ideal solution to the toy problem would be robust to increases in the power of the simulated humans, ideally to arbitrarily large increases.

There may be problems during training

An AI system may make mistakes during training. Being a "good" learner means that those errors will be quickly corrected, not that they will never occur. We would like to build systems that are robust to these errors.

Most of these mistakes will be innocuous, but theoretically they could be quite bad. For example, amongst our space of possible models we might imagine that there is one which tries to solve the problem and one which tries to survive (kind of like life on Earth). The latter model may instrumentally give good answers to questions, as long as that's the best way to survive, but might suddenly behave strangely at an inopportune moment.

I don't think this particular scenario is likely, but I do think that we should strongly prefer schemes that are robust to adversarial errors. I think that a scheme that can't handle this kind of error has a significant chance of suffering from other serious problems. Being robust to adversarial errors essentially requires exerting "positive pressure" on our models to do exactly what we want, rather than trusting that if they have mostly done what we want so far they will continue to behave as intended.

In the context of the toy problem, we might imagine that there are two different humans in the simulation, and that in each step we interact with one of them at random. If one of the two has been doing much better so far, we gradually shift our probabilities towards consulting that one exclusively. One of these humans is really trying to maximize their own reward, but their evil twin has an ulterior motive (like influencing our world in some way—which instrumentally leads them to achieve high reward, so that they can keep playing the game).

An ideal solution to the toy problem would continue to work in the presence of this evil twin.