Against mimicry

Paul Christiano Follow
Sep 19, 2015 · 4 min read

One simple and apparently safe AI system is a "copycat:" an agent that predicts what its user would do in a situation and simply does that. I think that this approach to AI safety (and AI) is more sensible than it at first appears, but that it is ultimately seriously flawed.

Initial objections

There are a few obvious problems with this proposal, which turn out not to seem too serious.

- What does it mean for me to be "in the AI's situation"?

We can consider a user operating an AI by remote control. That is, the user continues to do things that are in their interest, but their outputs have exactly the same format as the AI's outputs.

The human has additional inputs that the AI lacks, but these will be naturally absorbed into the prediction problem being solved by the AI, i.e. the predictor can try to predict the human's behavior given the sensor readings available to it.

- The human may not be that good at the task in question.

Even today, robotic control systems routinely perform tasks that would be very difficult for a human controlling a robot manually. In the future we can expect the problem to get worse as AI systems acquire additional capabilities that humans lack.

I suspect that this problem can be addressed by having the AI imitate a (human+AI) system. That is, the human user can get help from one AI (e.g. one that performs basic motions) in order to produce the training data that will be used to train another AI (e.g. one that strings together basic motions in order to perform a complex task). See here and here for a more extensive discussion of these ideas.

- Predicting what a human would do is a very roundabout way to solve the problems that humans are solving.

It's worth noting that more typical AI techniques can be used as tools to help predict human behavior: I discuss this here. For example, if you want to predict what plan a human will follow, a planning algorithm might help—we need not restrict our attention to what we normally think of as "prediction" algorithms.

Advantages

The main advantage of the copycat, from a safety perspective, is that it "never does anything you wouldn't do." That cleanly rules out most of the possible catastrophic outcomes. For example, you can be quite confident that the AI won't hatch an elaborate scheme to take over the world—unless that's what you would do in its place. Moreover, these failure modes are ruled out in a way that is exceptionally simple, robust, and easy to understand.

A big problem

Humans and machines have very different capabilities. Even a machine which is superhuman in many respects may be completely unable to match human performance in others. In particular, most realistic systems will be unable to exactly mimic human performance in any rich domain.

In light of this, it's not clear what mimicry even means, and a naive definition won't do what we want.

For example, suppose that a human is trying train a robot to stack some blocks. Suppose that the human does one of two things in any given trial:

1. Successfully stacks the blocks. This happens 95% of the time and takes about 3 seconds.
2. Fails to stack the blocks—the pile collapses after adding the second block.

Consider a robot which is able to stack the blocks only by proceeding more slowly, taking 10 seconds. However, the robot is fine at failing, which it can do in a way nearly indistinguishable from the human.

For many natural definitions of "best," the "best" way to mimic the human is for the robot to fail 100% of the time. For example, this will maximize the probability that the robot and human do the same thing; it will minimize the KL divergence between the distribution of

the robot's and the human's behaviors; it will maximize the difficult of distinguishing the robot and the human's behavior.

Indeed, what we really want—the "right" emulation—is a thing that the human never does: to stack the blocks successfully and slowly. This seems to be a fundamental problem with the copycat strategy, which can't be resolved in an easy and principled way.

Conclusion

Of course we can still train an algorithm to imitate a person, but we don't actually have any guarantee at all like "the agent won't do anything I wouldn't do," and moreover we don't even want such a guarantee. The outcome will depend on what our system is actually doing—on how it approximates an objective that it cannot exactly meet (or even meet any quantitative approximation to).

The most obvious response is to use maximization, even if our goal is to imitate human behaviors. We could apply this maximization either at the level of outcomes (e.g. inverse reinforcement learning) or at the level of actions (e.g. maximizing a rater's score for how well those actions reproduce the intended behavior).

I think falling back to maximization is a sensible response, and the only one which we really understand at the moment. But I'm also interested in avoiding some of the possible problems associated with maximization, and so I think it's worth spending some time trying to "fix" direct imitation.