Safe AI from question-answering

Paul Christiano Follow
Apr 10, 2015 · 5 min read

(Warning: minimal new content. Just a clearer framing.)

Suppose that I have a question-answering box. You give the box a question, and it gives you an answer. For simplicity, let's focus on yes/no questions, and assume it just outputs yes or no. Can we use this box to implement an efficient autonomous agent that is aligned with human values?

Let's make the following assumptions:

1. The box never gives an answer predictably worse than the answer that some reference human would give. That is, there is no distribution over questions Q where we expect the human to guess the right answer more often than the box.
2. Whatever language the box uses, it's *at least* as expressive as mathematical notation.

Under assumption [2], the worst possible box is a "math intuition module," a box which takes as input a mathematical assertion and outputs a probability. We'll assume that we are working with such a box, though the whole thing would just be easier if we had a box which could answer questions in e.g. natural language.

A proposal

We'll describe an AI which outputs binary actions. To make a decision, it asks the following question:

> *If the user considered my situation at* ***great*** *length, which decision would they think was best?*

It then takes the most probable action.

Formally (with links):

- Step 1: We provide a precise definition of "the user" as an input-output mapping, by giving concrete examples and specifying how to learn the model from these examples. This specification can't be used to efficiently find a model of the user, but it does give a formal definition.

- Step 2: We provide a precise definition of "considered my situation at **great** length" by defining a simple interactive environment in which the user can deliberate. The deliberation may be very, very long—it may involve growing smarter, simulating an entire civilization, developing new branches of science, examining each possible decision's consequences in detail… The interactive environment is designed to facilitate such a prolonged deliberation.
- Step 3: The user eventually outputs either 0 or 1. The choice may be made based on a deep understanding of what the AI is thinking and why, and what the effect of a particular strategy would be on the AI's decision. This seems like a good idea for getting a good outcome, but it also leads to a complicated analysis. For the purpose of simplicity, we'll assume that the after deliberation the user represents the quality of each possible outcome by a utility in [0,1], and then chooses an action randomly so that the difference in the probabilities is equal to the difference in the utilities that would result if the AI took one action or the other. (Note that this depends not only on the state of the world, but on the behavior of the AI in the future)

Given the simple interpretation of step 3, our AI is maximizing the utility function defined by the user's judgment-after-deliberation. But the user has a great deal of latitude here, and they could e.g. direct the AI to make arbitrary changes to its own code. So this procedure doesn't commit the user to bounded utilities, or to essentially any non-trivial philosophical or normative positions.

Note that any interesting behavior for this agent will be composed of a long sequence of binary decisions. But we can express any message in binary, so this is a weak assumption.

This procedure can be easily adapted to AI that maintains state, by including the AI's current state in the question: ("What would the user want the AI to do, when its state is given by X?"). We can use similar questions to find appropriate updates to the AI's state based on its recent observations. The deliberation then needs to consider a weighted average over states of the world that correspond to a particular state for the AI. We'll set aside this complication for now.

A proof

Theorem. Let A be the AI described above, and let B be another AI. Assume that the user follows the simplified strategy described in step 3 above. Suppose that running B results in a higher expected utility, according to the user's judgment-after-deliberation, than running A. Then, by using B we can sample from a distribution over questions for which we expect B's answer to be right with higher probability than A's answers.

So in some sense, this design makes optimal use of the question-answering system we have available. We can't get a better outcome without also building a better question-answerer—at least not if the deliberative process successfully captures the user's values. So the quality of the proposal comes down to the quality of the deliberative process and the model of the user.

Proof. To produce the distribution, we simply run B and choose a random question it is posed. It's easy to evaluate the sum of the differences between P(B is right) and P(A is right) over this whole sequence: it's a telescoping sum that adds up to U(taking B's decisions) - U(taking A's decisions). So if B receives a higher utility, B also has a higher probability of being right for a random question from this distribution.

A problem

I discuss some possible problems in the posts linked under step 1 and step 2 above, and in this post describing an older version of the proposal. There seems to be one major problem (and many minor problems), with several facets:

- The quality of the scheme rests on the quality of the user's outputs upon deliberation. But we can never actually run this deliberative process, and so we can't test if our definition is reasonable.
- During the deliberative process, the user's experiences are completely unlike their everyday experiences. This may introduce some divergence between the actual user and the hypothetical user. For example, the hypothetical user may become convinced that they are in a simulation and then start behaving erratically.
- Because we can't ever run this deliberative process, we can't ever give our AI training data for the kinds of problems that we want it to actually solve. So we are relying on a kind of transfer learning that may or may not occur in practice.

- The last point raises a general concern with the question-answering system as a model of AI capabilities. It may be that this just isn't the kind of AI that we are likely to get. For example, many approaches to AI are based heavily on finding and exploiting strategies that work well empirically.

I don't think these problems are deal-breakers, though I agree they are troubling. I think that if the system failed, it would probably fail innocuously (by simply not working). Moreover, I think it is more likely to succeed than to fail malignantly.

But taken together, these objections provide a motive to look for solutions where we can train the AI on the same kind of problem that we ultimately want it to perform well on. We can capture this goal by trying to build a safe AI using capabilities like supervised learning or reinforcement learning.