Universality and model-based RL

Paul Christiano Follow
Jan 9 · 9 min read

Ascription universality seems sufficient for informed oversight, and I've argued it could be used to prevent cascading failures in HCH. In this post I'll argue that ascription universality also appears to address two problems with aligning model-based RL:

1. It allows you to perform induction without being hijacked by malicious predictors (this is the "benign induction" problem, which is extremely speculative)
2. It seems necessary, though maybe not sufficient, for extracting "all the relevant facts a model knows." (Which is at least necessary, and I think probably sufficient, for defining a reward function that yields aligned behavior when optimized.)

These arguments aren't rigorous enough that I find them convincing, but they are enough to convince me that better understanding universality is a top priority. At this point I wouldn't be surprised if universality is enough to resolve most of the classical philosophical difficulties for AI safety. If so, first we'll need to better understand in what sense it could be achievable.

Setup

I'm interested in the following hopefully-aligned version of model-based RL:

- Use iterated amplification to learn a (distribution over) models of the world.
- Use iterated amplification to learn a utility function over sequences of states and actions.
- Plan in that model+utility function (e.g. using MCTS)
  (You might want to combine learning a value function and terminal values, but for now I'll simplify.)

This is similar to the traditional model-based RL approach, except that we learn a model and utility function with iterated amplification,

instead of learning a model to predict future observations and defining a utility function over observations by hand.

What is a model?

The simplest kind of model to consider formally is a *predictor*, which we can view as a probability distribution over sequences of observations. We can evaluate predictors on past observations, and use them to predict future observations.

This is not enough for making the decisions, because the utility of an outcome is not a (simple) function of our future observations. We care about lots of features of the world that are hard to directly observe, and if we tried to define our preferences in terms of our observations alone we'd need an additional step where we map our observations to a prediction about what's "really happening" out there in the world. For example, given a bunch of observations we might ask "given those observations, is Alice actually happy?" At some point we need to get a a distribution over models that we can use to answer questions about *the stuff we care about*, and not just our observations.

(Closely related: Formalizing Two Problems of Realistic World Models, though the basic philosophical issue is a classic.)

As a first step we can ask a model to answer arbitrary questions about the world rather than only reproducing observations. We can evaluate a model by looking at questions whose answers we already know. A problem with this is that in order to behave well a model only needs to correctly answer *the kinds of questions that we might know the answer to*, and the simplest model may well behave strangely on other questions.

A second step is to view a model M as an object that HCH can reason about, and use HCH to answer questions like "If model M describes reality, then what would be true?" For example, a model might posit some kinds of physical objects, and HCH could reason from those objects both to explain our observations and to infer the existence of other objects we care about. This seems to more closely track the reasoning humans perform about possible models, and (a) gives us more traction on a good prior over models, (b) gives us traction to pose other questions about the model (beyond what its predictions are).

I'll use this notion of "model" throughout the post.

Aside: defining goals

If we are able to learn a "correct" and "understandable" model, I'm optimistic about being able to find a utility function that can induce aligned behavior. In particular, I think that function can look something like:

- Identify a thing that is "me" in the resulting state
- Conservatively evaluate how many resources "I" effectively control
  (working together with AI systems, *etc.*)
- Conservatively evaluate how "I" changed over the history—do I endorse the way in which my beliefs and values shifted?

We can't hope to have completely robust answers to any of these questions. But we can at least hope to have a utility function that is safe to maximize within our model, by being aware of the history leading to high utility and penalizing the act if that history leads to high utility for bad reasons.

For example, if we can see that we control resources only because other agents disagree with our assessment of which resources are important, then we can revise our estimates appropriately—we don't have to fix a single way of valuing resources which is robust to arbitrary adversarial pressure.

(Realizing this kind of hope is related to the ideas expressed by Jessica Taylor here and here.)

As another example, our preferences over changes to ourselves will not be completely robust. But maximization over the utility function will only exert pressure for such changes if they somehow lead to *higher* reward. And in order to do that, the changed behavior needs to actually manifest over the history, and we can use that manifestation to detect a corruption that allows the agent to receive higher reward even if we wouldn't recognize most possible corruptions.

"Knowing what the model knows" seems to be a critical ingredient for constructing such a utility function. It's not clear whether or not it's sufficient, but I won't discuss this issue more in this post.

Problem 1: Benign induction

If we take the "simplest" model that explains our observations, it could potentially be very strange. In particular, it could produce correct observations because it contains an intelligent predictor rather than because it captures "real" physics. This could cause it to generalize in unpredictable and dangerous ways (depending on the motivations of that agent).

Two variants of this problem have been discussed:

1. The predictor may literally *contain* intelligent agents who are trying to predict our observations. These agents may be motivated by the desire to survive the training process, but then may generalize maliciously once off the training distribution.
2. The predictor could *reason about* intelligent agents who change their behavior in order to explain our behavior, because the existence of such agents is the "simplest" explanation for our observations. A notion of simplicity that produces this behavior is in some sense not the "right" one, but I believe it would probably apply to the simple formalisms we currently understand (such as minimum description length).

(A single situation may have some features of both #1 and #2, this is not a clean dichotomy.)

If we have a universal reasoning procedure, I think we can likely avoid these problems.

Option 1: filter malign models

We use the following procedure:

- Compute log likelihoods for models as normal, by asking HCH("how likely are facts F if model M is true?")
- Additionally, compute HCH("is model M likely to lead to catastrophically bad predictions on inputs we'll encounter in the future?") and judge a model "unsafe" is the answer is "yes."
- Apply Bayes as usual but restricted to the safe models.

Why might this work?

If a model M is malign, then it is making good predictions *because* of the existence of some adversarial agent. Then:

- If that agent is actually contained in the model (situation #1), then universality means we can learn everything that agent knows. But the agent has formed the plan of giving good answers in order to increase the model's log likelihood and so have more influence later, and if we know this then we will realize the model will generalize poorly and filter it out.
- If the agent is being reasoned about by the model (situation #2) then things are even easier, since the model knows about the existence of the agent and its motivation. Then applying universality to the model itself will let us know the model would generalize poorly.

Going beyond the simplest safety test, we could also try to throw out any "abnormal" models, e.g. where the stars don't actually exist or our local universe is actually a fiction. This could leave value on the table if these weird hypotheses are important, but it could also reduce the risk of making a serious error based on strange hypotheses. Universality is relevant because it potentially allows us to avoid adversarial pressure on this definition of "normal"—if a model looks normal because an adversary is trying to make it look normal, then we can recognize that and throw it out as unsafe.

Option 2: outcompete malign models

Suppose that I have a prior μ over models, and I believe that the posterior puts significant probability on "malicious" models that make good predictions only because they contain agents who are motivated to make good predictions. (Or because they reason about such agents.)

This is a strange epistemic situation to be in. I must believe that on average these malicious agents have a "better" prior μ than I do—that's what they need in order to assign a higher probability to the data than the rest of μ. Jessica Taylor writes about this situation here, and has similar intuitions about it.

Intuitively, something like universality should allow you to avoid this state. That is, the malicious models must perform some reasoning that we believe allows them to assign better prior probabilities; by ascription universality HCH must be able to capture this reasoning and use it to inform *its* prior. As long as malicious agents have a reasonably low prior they will therefore end up with a reasonably low posterior.

To implement this strategy, we could use the same procedure as in the last section, but instead of using a simplicity prior over models M, we would use HCH to compute those prior probabilities by asking HCH("a priori how likely is model M to describe the observations we are going to provide to it?") or something along those lines.

I don't know whether this works, and I'm not sure it will really be possible to say much precisely without having a cleaner picture of the sense in which HCH is or isn't universal. If it does work, this seems like a much cleaner solution than filtering. The fact that there are two plausible options on the table makes me more optimistic that universality will address benign induction one way or another.

Problem 2: extracting facts from a model

Consider a model M that specifies low-level physical principles and some basic "bridging" facts about how everyday objects are implemented in low-level physics. HCH can answer questions about the world by performing a detailed physical calculation, and then using bridging facts to relate the resulting state to things we care about.

(Modeling physics itself is a somewhat exotic example, but I think the same phenomenon would occur in more prosaic cases, wherever there are effective but opaque models.)

Unfortunately, M can be a good model even if the bridging facts are not exhaustive. They might allow HCH to answer some kinds of questions—e.g. to predict the observations that we're using to test models—but not to answer other important questions.

For example, it may be the case that M predicts that Alice will report high satisfaction with the AI's behavior, by performing some physics calculation and then using bridging facts that relate physical states to Alice's reports. At the same time, the physics calculation might actually simulate Alice being bribed to report high satisfaction, but it might not be possible to infer that from the available bridging facts.

But the physics computation contains a complete simulation of Alice being bribed, and so we can ascribe knowledge of the bribery to it. Thus if HCH is ascription universal, we can hope it recovers all of these facts.

It's not obvious whether universality is actually sufficient for this purpose, and it will depend on exactly what form of universality ends up being achievable. It might be that a "reasonable" ascription strategy would recognize that the model knows facts about atoms, but not facts about Alice.

At a minimum universality is *necessary* for this purpose—if M can believe facts that HCH can't identify, then those facts could turn out to be relevant to the utility evaluation. It's just that there might or might not be a further step not captured by universality.

My current guess is that the most natural form of universality will give us everything we want, and that thinking about "ontology identification" is mostly useful for finding hard cases for universality. Regardless of whether that's true, I think the next step should be a clearer picture of prospects for universality, after which we can revisit this question.