Informed oversight

(revisited)

Paul Christiano Follow
Jan 9 · 8 min read

To train an agent with RL, we need to answer the question: "which of the trajectories τ^1 or τ^2 is better?"

Likewise, in order to train for acceptable worst-case performance, we probably need to answer a question like: "is trajectory τ acceptable?"

We want to answer these questions "well enough" that our agents end up aiming for good outcomes on average and always behaving acceptably. But how well is well enough?

In this post I'll argue that if our goal is to train a well-motivated agent, it is necessary and sufficient for our answers to reflect *everything our agent knows*. This seems essentially equivalent to ascription universality of the oversight process.

(This post is an update of these 2016 posts. My understanding of the problem has improved a lot over the last few years.)

Information vs incentives

We can distinguish two functions of feedback:

- It provides *information* to the agent we are training. For example, if I provide negative feedback when an agent serves me pizza, it can learn that I don't like pizza. (This might happen by gradient descent selecting for agents with different beliefs, or by an RL algorithm that uses past reward signals as information.)
- It determines the *incentives* for the agent we are training, i.e. is used to select the motivations of the agent. For example, an agent pre-trained to make predictions about the world might already understand that I dislike pizza. If I provide negative feedback when the agent serves me pizza, then gradient descent will select for agents that are motivated to avoid things I dislike (at least instrumentally).

From the perspective of information, there is no such thing as "good enough." More information is always better.

This post is about incentives. We can aim for a "correctly motivated" agent, who lacks relevant information but is at least trying to do the right thing. At that point there is nothing to gain from better oversight.

So our question is: how well do we have to answer these questions to produce an approximately correctly motivated agent?

(In the long run, I think that incentives are the more important function of feedback. Agents should mostly learn about the world from other richer signals. So I'm not worried about whether our feedback is optimal from an information perspective.)

Necessity of knowing what the agent knows

Suppose that our agent understands facts about the world that we don't understand, i.e which aren't reflected in our answer to "which of the trajectories τ^1 or τ^2 is better?" I'll argue that this is a bad situation.

If the agent literally took actions and made observations that weren't available to us, then we would clearly have a problem. For example, consider an agent trying to defend a computer system from attackers. If the agent has access to actions or information we are missing, then it might deliberately take actions that compromise the tools we would use to determine whether our system has been compromised. Because we can't observe this malicious behavior, we can't assign it a low reward. And after compromising our tools well enough, the agent could ensure that they always paint a rosy picture of the situation regardless of what actually happens. We intended to protect ourselves from attack, but instead introduced a powerful attacker with privileged access to our system.

The same thing can happen if we observe everything our agent observes, if we aren't able to understand everything our agent understands. In the security example, literally seeing a sequence of bits moving across an interface gives you almost no information—something can look innocuous, but cause a huge amount of trouble. In order to incentivize our agent to avoid causing trouble, we need to be able to detect any trouble that the agent deliberately causes. Even an apparently mundane gap in our understanding could hide attacks,

just as effectively as if we'd been literally unable to observe the agent's behavior.

We might hope that an attack will ultimately cash out as *something* that we can unambiguously observe, so that this is merely a problem with short time horizons. But I don't think that's the case—ultimately we are always restricted to indirect observations of the things we care about. We rely on second-hand reports, on viewing small parts of a situation that we hope are representative, on cameras and other sensors that purport to describe the world. All of these things would also be manipulated by an agent incentivized to make things look good to us, and understanding what the agent understands is the most natural defense against that kind of manipulation.

Intuitive argument of sufficiency

Suppose we "know everything that the agent knows." That means the agent should defer to our estimate for any unknown quantity.

Now consider two possible motives an agent could have:

- Create trajectories that are actually good, according to our "real" preferences (extrapolated as we wish that extrapolated…)
- Create trajectories we believe are good

If the agent defers to us about the goodness of trajectories, then these two motives will lead to identical actions. So selecting for agents that take actions that we think are good also optimally selects for agents actually motivated to do good.

Intuitively, imagine a student who is trying to predict whether their teacher will think a statement is true. If the teacher knows everything the student knows and can see every consideration the student sees, then the student shouldn't expect to be able to predict any systematic error in the teacher's beliefs. So trying to predict what the teacher believes is exactly the same as trying to predict what's true.

Tightening the argument

The argument in the last section isn't quite right. Most glaringly, if the overseer is universal then it means *we* think the overseer has better beliefs than the agent, but it doesn't mean *the agent* thinks that.

Changing the statement. To fix the argument, we need to be a bit more careful. We'll aim to show that if we think the overseer is ascription universal, then:

- We expect the agent A to achieve at least as much utility (by our real, unobserved utility function) than any action that can be easily computed using any set of beliefs that can be ascribed to A —and in particular A achieves at least as much utility as if it were perfectly motivated.

We can't infer that A is perfectly motivated, because it may have arbitrary beliefs that differ from those of the overseer. For example, A might believe that the world will certainly end in a week; instead of planning under that assumption (as it would if it were "correctly" motivated), it will learn to choose actions that the overseer prefers. From its perspective it is choosing actions suboptimally in order to please the overseer, but from our perspective this is better than if it were "correctly" motivated.

Another way to see we need this weakening: many different believes can be ascribed to the agent, and it can't simultaneously behave as if it were perfectly motivated with respect to each of those sets of beliefs.

Assumptions and proof. This will mostly just be notation; other than that it follows almost immediately from the definition of ascription universality, though we need to do a little bit of work to massage the definition into the right form.

Let $\mathbb{E}$ be any subjective expectation w.r.t. which the overseer is ascription universal.

Let $\mathbb{E}^O$ be the expectations of the overseer. Let U be our actual utility —this is not a well-defined function we can actually evaluate, but we can query a subjective expectation about it. Then the reward function is $r(a) = \mathbb{E}^O[U(a)]$. Note that the overseer doesn't condition on the fact that the agent produced this particular action (so no one needs to reason about the training process itself).

Let $\mathbb{E}^A$ be some expectations that can be ascribed to the agent. Let A be the expression representing the agent's output (**after training**), and let a be the actual output (we need to distinguish the symbol from the action itself because we are plugging this into a subjective expectation). Let A′ be a particular different definition of an action

the agent could feasibly output, e.g. argmax $\mathbb{E}^A[\mathbb{E}^O[U(\cdot)]]$, and let a' be its actual value.

Because A and A′ can both be computed using the beliefs $\mathbb{E}^A$, we can apply the epistemic dominance property to predicates defined in terms of A, A′. (There is the subtlety that dominance requires properties to be open in the product topology, which can be avoided by carefully considering randomization and only becomes relevant when the utility function U itself depends on the behavior of the agent.)

We'll assume that $\mathbb{E}$ believes that training converges well enough that the action A will receive a higher expected reward than A′ (otherwise the agent would eventually switch to outputting A′). We also assume that $\mathbb{E}$ knows that the overseer epistemically dominates the agent, and that it satisfies a standard iterated expectation property.

Then we have:

- $\mathbb{E}[a \leftarrow A; a' \leftarrow A'; U(a) - U(a')]$
  $= \mathbb{E}[a \leftarrow A; a' \leftarrow A'; \mathbb{E}[U(a) - U(a') \mid a{=}A, a'{=}A']]$ (*iterated expectation*)
  $= \mathbb{E}[a \leftarrow A; a' \leftarrow A'; \mathbb{E}[\mathbb{E}^O[U(a) - U(a')] \mid a{=}A, a'{=}A']]$ (*dominance*)
  $= \mathbb{E}[a \leftarrow A; a' \leftarrow A'; \mathbb{E}^O[U(a) - U(a')]]$ (*iterated expectation*)
  $= \mathbb{E}[a \leftarrow A; a' \leftarrow A'; r(a) - r(a')]]$ (*definition of reward*)
  ≥ 0 (*training is converged*)

As desired.

(The meaning of the $\mathbb{E}[\ldots;\ldots]$ notation is to take the expected value of the output of the program which performs the sequence of steps separated by ;'s, returning the final value.)

This argument only constrains the equilibrium, but eventually we need to say something about the entire training process. It looks to me like very similar ideas are needed for the more general case, but it's a lot harder to make the arguments precise.

Corrigibility

So far we've discussed reward functions that induce correct motivations. The other half of our problem was recognizing unacceptably bad behavior: mistakes so bad that they cannot be corrected.

Again, we can't hope to recognize all such mistakes. For example, it may turn out that a particular physics experiment will destroy the world, but correctly predicting that would require a detailed understanding of physics.

But if we know everything the agent knows, then we can avoid cases where the agent intentionally brings about a catastrophic outcome. In particular, we can avoid cases where the agent takes an action precisely because it would prevent us from correcting the agent's behavior or reasserting control over the situation.

Ascription universality seems to be just what we need for this purpose as well: if we believe the agent is pursuing a plan *because* it prevents us from correcting the situation, then we can ascribe that belief to the agent, and so an ascription-universal overseer will recognize that the behavior is incorrigible and can deem the behavior as unacceptable.

Such an overseer may also sometimes flag as unacceptable certain actions that are catastrophic but which the agent didn't recognize as catastrophic. But I think that's fine, I'm not unhappy if we end up training the agent to avoid a catastrophic mistake.

What does this actually look like?

If we are training a model with RL, my best candidate for obtaining a universal overseer is using iterated amplification to train a second head on the agent to answer relevant questions, including questions of the form "Which of these two trajectories is better?"

There are difficulties from training a model to play a game where it shares almost all of its activations with its opponent, and with ensuring that the overseer can remain sufficiently competent relative to the agent (without slowing down training). These resemble the discussion of the `Info` proposal here, but rather than having a separate overseer who evaluates how useful information was, we directly optimize the side information by using the same amplification process. Both of these issues seem OK with gradient descent, but they are larger problems for some other search algorithms, so I don't consider the issue settled.

This proposal also faces all of the usual difficulties of iterated amplification—it's not clear how to generate the distribution of questions, whether the training process is stable, whether errors compound, whether the steps of amplification increase the

complexity of training, and so on. These issues seem practically important, and keeping informed oversight in mind can help us understand exactly where the bar is for an adequate solution. I tentatively think the conceptual difficulties in achieving ascription universality are more likely to be a fundamental obstruction.