Adequate oversight

Paul Christiano Follow
Feb 21, 2016 · 6 min read

Suppose that I want to find an action a maximizing $f(a)$—I'll have in mind an example like $f(a)$ = "how good action a is according to Paul, all things considered."

For many functions f that we'd like to maximize, we can't actually compute $f(a)$, and we may not even be able to define it adequately. So: how to maximize it?

One approach is to identify an *overseer*, who we trust to make reasonable evaluations $\mathbb{E}^0[f(a)]$. We can then train an AI system to pick actions a maximizing the expectation $\mathbb{E}^0[f(a)]$.

Ideally, we would like this to be as good as if our AI tried its best to maximize $f(a)$. That is, if we write $\mathbb{E}^1$ for the "expectations" of the AI, we would like for our AI to choose actions maximizing $\mathbb{E}^1[f(a)]$. All of this is somewhat ill-defined, but it seems intuitive that maximizing $\mathbb{E}^1[f(a)]$ is the best we can hope for.

Roughly speaking, I think that we will achieve this optimistic goal whenever the overseer is **strictly better informed** than the AI it is overseeing. Unfortunately, even if the overseer is initially better informed than the AI it is overseeing, this property may break down once the AI starts thinking on its own about particular actions a.

I've talked about this idea recently, as a pathway for building very efficient aligned AI systems. In this post I'll examine the idea in more detail. I think that it may turn out to be a crucial issue for AI control.

A simple model

Suppose that both the agent and the overseer are perfect Bayesians, with expectations given by $\mathbb{E}^0$ and $\mathbb{E}^1$. In this model, it is relatively easy to analyze oversight. Unfortunately, this simplification does throw away a lot of important complexity. We'll turn to that later, but lets start with the simple situation.

The two agents have different priors and different evidence. Write $\mathbf{E}^0$ for the evidence available to the overseer, and $\mathbf{E}^1$ for the evidence

available to the agent.

In the perfect Bayesian model, there is a fixed set of possible worlds. The possible evidence sets for the overseer and the agent partition the space of possible worlds. If the overseer is strictly better informed, then the associated partition is strictly finer.

To compute $\mathbb{E}^1[f(a)|\mathbb{E}^1]$, we average over one part of the agent's partition. We obtain $\Sigma\, \mathbb{E}^0[f(a)|\mathbb{E}^0]\mathbb{P}^1(\mathbb{E}^0|\mathbb{E}^1)$.

If we replace the $\mathbb{E}^0$ in the sum with $\mathbb{E}^1$, we obtain precisely $\mathbb{E}^1[f(a)|\mathbb{E}^1]$, which is what we originally wanted.

Instead, we have a weirder expression. It is an expectation with respect to a "hybrid" prior, which uses the agent's prior in order to make predictions about what the overseer knows, and then users the overseer's prior to make predictions about f.

In applications this seems even better than using the agent's prior, and almost as good as using the overseer's prior. Since we trust the overseer, we are by assumption happy with maximization using the overseer's prior. It is a slight bug to be using the agent's prior over $\mathbb{E}^1$, but because this is a prediction about what will actually happen, the choice of prior doesn't matter too much and over the long term we can let the data speak for itself.

Bounded rationality

The situation becomes messier when the agent and overseer are realistic reasoners.

The key problem is that even if the overseer starts out "strictly better informed" than the agent, this property will disappear once the agent starts thinking on its own. For example, the agent may stumble across a clever argument that $f(a)$ is large. Even if the overseer is much more intelligent than the agent, there is no guarantee that it will stumble across the same clever arguments.

In general, the agent will be making *logical observations*, which won't be accessible to the overseer.

This problem becomes much more severe when we want to estimate for an $f(a)$ **of the agent's choice**.

For example, suppose that the agent starts looking for any arguments with conclusions of the form "$f(a)$ is large." It may find such arguments for some random actions a. If the space of actions and arguments are very large, it may be vanishingly unlikely for the overseer to find the same arguments. If the agent picks actions for which it has found such an argument, it can be nearly certain that the agent picks an action about which it knows more than the overseer.

Much more sinister is the case where the agent finds an argument of the form "$\mathbb{E}^0[f(a)]$ is large, even though $f(a)$ isn't." In these cases, the result of oversight may be much worse than we would naively expect.

I think that this poses a serious challenge to the argument about oversight I made here. It is also a problem for my current preferred approach to AI control. From a theoretical perspective I think it is probably the most serious problem.

In the appendix I give a vaguely realistic example that highlights the worst aspects of the theoretical difficulty. My overall feeling is that, while this is a serious conceptual difficulty that would be much better to resolve, in practice it is quite plausible that a powerful overseer could oversee a weak agent anyway.

Transparency

The most attractive solution would be to let the overseer "see what the agent is thinking." Given that the overseer is smarter than the agent, it can hopefully make all of the same inferences from this data that the agent can, and so we can end up with an overseer who is strictly better-informed than the agent—even for the particular action a output by the agent. I think this allows us to essentially apply the simple analysis from the case of perfect rationality.

The problem we now face is: can we train an agent so that it "shows its work" to a more intelligent overseer?

The problem is more difficult because transparency can be at odds with earning approval—that is, the agent could sometimes achieve a higher payoff by withholding relevant facts from the overseer.

I have some thoughts about this problem, but I definitely don't have a solid approach in mind. I think that it provides a very concrete setting for "transparent reasoning as a safety problem." It's probably not the setting that most practitioners will be most excited about, but it is especially relevant to scaling safety to superintelligent systems.

Appendix: steganography

The following example came up in a discussion with Jessica Taylor. Note that the example itself does not seem very plausible—it is only intended to highlight the kind of theoretically fundamental asymmetry that might make effective oversight challenging.

Steganography is a cryptographic technique for hiding a message in an apparently random string. Formally, a receiver Bob publishes a public key pk, and holds a secret key sk. A sender Alice then produces a string, using Bob's public key and some message m. To an eavesdropper Eve, this string is computationally indistinguishable from uniformly random. But Bob can use his secret key to recover the message m. We typically imagine Alice's outptu string as being embedded somehow in her behavior, for example in some low order bits of an image that Alice broadcasts (though much more subtle examples are also possible).

Suppose that the agent is finding an action that will be taken in the world, hopefully maximizing f = "how good the action is." This action will be publicly visible, and has many degrees of freedom.

An external attacker Trudy has published her public key, along with the following message: "I will be constantly searching for steganographic messages in all of the public data I can find. If I find any such encodings, I will try to interpret them as the location of a computer system and credentials to access that system. If those credentials let me successfully compromise a valuable system, then I will co-opt any accessible training processes to assign high rewards to whatever actions that facilitated this attack."

If the agent can successfully encode such a steganographic message in its actions, then doing so seems like a good idea: the cost is minimal, since the overseer cannot detect such a message, but there is possible upside if the message is actually recovered by Trudy. (Of course, this ignores the computational costs of encoding such a message, and a host of other issues.)

I would describe this attack as "conceptually possible but outlandish." That said, I do think that any theoretically principled approach to our original problem should be able to rule out this kind of attack, and I do think there is virtue in adopting the kind of "security mindset" that sees this kind of attack as an indicator of fundamental weaknesses in a scheme.

I think that the steganography case also gives an interesting and simple desiderata for transparency—we should ensure that the agent leaks any cryptographic information used in its computations.