AI control on the cheap

Paul Christiano Follow
Jan 31, 2016 · 7 min read

Ideally, we will build aligned AI systems without sacrificing any efficiency.

I think this is a realistic goal: I'm aiming for modest *additive* overhead, which does not depend on the complexity of the problem or the capability of our AI.

Several commenters have expressed skepticism about this goal, and especially about whether the methods I am describing could possibly achieve it.

In this post I'll argue that approval-directed agents (based on counterfactual oversight) might achieve additive overhead. I'm choosing this example for concreteness, but I don't really mean to advocate for approval-direction. I'm mostly making the general claim that aligned systems could be very efficient.

It might feel like additive overhead is a "best case" that we could only meet if everything goes perfectly, and that in reality we will almost certainly be far away from it. But I think that there is actually a sense in which we have some slack; I claim there is a broad range of "good enough" implementations that will be no less effective as their unaligned counterparts.

(Note: the optimism in this post is only justified if we can solve the informed oversight problem. The difficulty of this problem was not obvious when I originally wrote this post. If it proves impossible, I am much less sure about whether we can build efficient aligned AI.)

The basic picture

We want to optimize an implicitly defined objective for which we can't give feedback— namely, we want to optimize how well the system's behavior satisfies our preferences.

My recent proposals construct training data that approximately reflects this implicit objective. For example, we might have the user evaluate how good an action is, with the help of AI assistants, as a

proxy for how good it actually is. Or we might use our system's estimate of how well its action meets some symbolically defined goal, like "how good the user would judge your action to be upon extensive reflection and self-improvement."

I'll argue that training on these proxies can yield good behavior, and that calculating them may not increase the cost of training too much.

During operation

I'd like to compare the approval-directed agent to an idealized benchmark, the algorithm which (informally) "actually cares" about doing what we want. Under appropriate conditions, I claim that counterfactual oversight exactly matches this benchmark.

The process of counterfactual oversight may involve extensive deliberation, or may involve a costly process for reasoning symbolically, or something else expensive. But during operation, this doesn't matter at all: we only use counterfactual oversight to gather training data. During normal operation, the system can immediately act with no additional overhead.

So the only question is whether our training procedure leads to good behavior. Obviously this will depend on what kind of counterfactual oversight we apply—I'll try to offer a sufficient condition.

(I'm going to have to talk informally because we don't have the kind of theory that could make these statements precise. My intuition is that there is a backing theory and that it would be worth developing that theory. But even if not, I think this section stands as a purely informal argument.)

Our system is taking actions to maximize the score we give it, A. Meanwhile, we "really care" about some function V, reflecting the full impact of an action on our values.

Personifying our system for a second: if $\mathbb{E}[A(a) \mid E] = \mathbb{E}[V(a) \mid E]$ for every action a and any evidence E, then maximizing A is equivalent to maximizing V, i.e. they correspond to the same procedure.

This corresponds to a simple requirement: our oversight has to capture everything about what we value, *as well as the agent can tell*. As long as we meet this goal, we aren't losing anything by training on A instead of V. We have some slack, in that the process defining A can be much more complex than the process defining a, and so the

oversight actually has a lot of room to be much *better* informed than the agent.

In my preferred proposal, oversight involves several AI assistants which are as powerful as the system being overseen. And oversight can take much longer than the original agent spent thinking. For example, we might spend several hours evaluating a decision made in a minute, and base our decision on hundreds of queries to comparably-powerful assistants. We can also inspect the internal state of the system being overseen (which may or may not be helpful, depending on how transparent it is).

Under these conditions, it seems quite plausible for A to reflect everything that the agent can figure out about V. It's certainly not obvious that it does, but it looks like a realistic goal to shoot for.

In abstract approval-direction, A is defined symbolically and the oversight would come from some estimate of A(a). Again, assuming that we are willing to spend significantly more effort on computing A than on choosing a, it seems quite plausible that it could capture all considerations discoverable by the agent. Indeed, I would take this as a requirement for a successful approach to symbolic reasoning.

A reduction?

In fact, any reasonably effective approach to AI alignment could be used as source of training data in this way. That is, if we can build an aligned system that is 10% as efficient as its unaligned counterpart, we could use it as an overseer to train an efficient aligned system.

Training

Even if our system works well at test time, the process of counterfactual oversight will increase the cost of gathering training data. And gathering training data could be a significant part of the overall system cost.

I think that better managing this overhead is a promising research direction. But my best guess is that the additional burden can be made minimal. There is a lot to say on this topic, and much of the reason for optimism is that there are so many promising approaches (many of which are standard research problems in AI). I'll try to convey a few intuitions behind my optimism.

One important observation is that, whatever other data we might have used during training, we can continue to use that as extremely

informative "unlabelled" data. The amount of extra training data we need, involving expensive human feedback, is determined by the *additional* complexity of satisfying our preferences above and beyond the complexity of performing the underlying tasks.

Suppose that we are concerned that our agent, trained with our approval A, will do worse than a reinforcement learner trained with a cheap reward signal R (such as "how much money the agent earned," "whether the human pressed the reward button," or whatever).

The simplest approach is to use R for variance reduction (though this shouldn't be necessary if we are sufficiently good at semi-supervised learning). That is:

- In every trial, we provide a reward of R to the approval-directed agent.
- In an ε fraction of trials, we consult the overseer to obtain A, and provide an additional reward of (A−R) / ε.

This gives an unbiased estimator of approval. The only restriction on ε is our desire to control the variance of (A-R)/ε, so if R is a good proxy for approval then we can choose ε quite small. Intuitively, we can also choose ε very small if A-R is relatively easy to learn, though I have no idea how to actually express that idea rigorously.

We could also train a more sophisticated model A′ to predict A (with a quadratic loss function), using all of the available information (including R). In every trial we provide our agent the reward A′, and in an ε fraction of trials we provide the correction (A-A′)/ε. The complexity of training A′ depends on the difficulty of learning what we will approve of, and is unrelated to the difficulty of the underlying task. And we can take ε to be basically the RMSE of our predictor.

Another simple technique is to actively pick important examples to label, or even to generate new examples that are maximally informative. Going further, we could ask arbitrary questions chosen to reduce uncertainty about A as quickly as possible, which may not simply be evaluating examples.

Finally, note that from an information-theoretic perspective the agent needs extremely little training data. For example, with a good predictive model of reality, the agent can anticipated the result of the oversight process without it ever actually happening. In practice a strong agent might be able to learn very quickly and do rapid transfer

from one domain to another. The technical lower bounds on oversight frequency come from our desire to handle adversarial errors rather than from any information-theoretic or algorithmic argument.

It's certainly not obvious we can achieve this. But again, I think that having minimal additive overhead is a reasonable and realistic goal to shoot for.

Conclusion

I don't think that we actually need to meet the zero overhead limit—I suspect society has enough ability to coordinate, and there are enough safeguards in place, that it would be OK if aligned AI was 10% more expensive or maybe even twice as expensive.

But I think that *aiming* for no overhead, rather than accepting large productivity losses, is valuable.

This view is in some sense the mirror image of Eliezer's; he argues that, whether or not it is a good idea to build a "sovereign" AI that acts autonomously in pursuit of human values, we should at least think about the hard problem. Thinking about the hard problem forces us to hold our ideas to a high standard, and to notice problems that we might otherwise have erroneously swept under the rug.

Similarly, whether or not we need to build aligned systems that are perfectly efficient, we should at least *aim* to build such systems. Thinking about the hard problem forces us to hold our ideas to a high standard, and to notice problems that we might have erroneously swept under the rug. In practice, I suspect that many designs that look like they have some "minor" inefficiencies or limitations will end up being completely impractical.

At some point I might conclude that the ambitious goal is unachievable, or at least too hard to be worth focusing on. But for now, the situation actually looks pretty good to me, and I think there is a good chance that we will be able to achieve additive overhead.