Abstract approval-direction

Paul Christiano Follow
Nov 28, 2015 · 14 min read

Consider the following design for an agent, which I first described here:

Pick an action a to maximize V(a) := "the extent to which the human operator would consider a to be a good action, upon reflection." (To give a formal definition of V we need to give a formal definition of "the operator" and "upon reflection.")

In this post I want to compare this proposal to a similar goal-directed design, in which we formulate an appropriate utility function U and then build an agent that tries to maximize U.

Many of these points were raised in my original post, but the key advantages and concerns have become much clearer over the last year.

Advantages

Avoiding possible problems

Researchers interested in AI control have spent a lot of time thinking about philosophical questions related to rational agency; this set of questions is well-represented by MIRI's research agenda.

One motivation for this kind of research is the view that without it, we don't have any idea how to describe what we *want*. We don't even know what it means to "act rationally in pursuit of a goal," so how can we prove that a system reliably acts rationally in pursuit of *our* goals?

The approval-directed approach can potentially avoid dealing with any of these issues. Of course, to the extent that these questions are important for figuring out what to do, our AI systems would necessarily have to think about them or adopt provisional solutions. But if we get these questions wrong, it's not really clear what harm we do:

- We don't commit to a particular formulation of decision theory or a particular theory of logical counterfactuals. To the extent that there are recognizably repugnant consequences of our current best guesses about decision theory, V will predictably recommend against acting on those consequences.
- We don't commit to a particular formulation of consequentialism or a particular representation of values (e.g. we don't require a representation by a real-valued utility function, we don't have to think about infinities or small probabilities).
- We don't commit to a prior (even over logical facts). We need a procedure for maximizing a given function; such a procedure might implicitly be optimizing with respect to a "logical prior," but we aren't committing to that prior in a substantive way. For example, our agent won't try to ensure that future AI's use the same prior, but will instead defer to V about what logical prior future AI systems should use.
- We don't commit to any assumptions about how the agent is embedded in the world. The function V will, for example, recommend against any action that would compromise the internal integrity of our AI. Learning this relationship doesn't require any philosophical heavy-lifting.

If we accept these points, then it's not clear why we need to have a better understanding of rational agency per se. For example, if we could prove that our system was really choosing actions that maximize V, then we would have (roughly) as much confidence as if we could prove that it was really choosing actions that maximize an appropriate utility function U.

Corrigibility

We would like to build AI systems which are aware of their own limitations and which behave conservatively in light of those limitations. For example, we would like to build systems which respond to user instructions, even if they think that they have a better idea.

The desired *behavior* seems almost immediate in the approval-directed framework. As a simple example, consider an agent which is considering (in Eliezer's words) whether to "disassemble [the user's brain] to update [its] model in a Bayesian fashion on all possible

information that [the user's brain] contains." It doesn't take a genius to figure out that the human won't approve of this plan.

Of course the user *might* decide, upon reflection, that actually the AI is pretty perfect and so it should just go ahead and disassemble their brain. But they should only approve if they not only believe that the AI is pretty perfect, but that the AI actually has enough justified confidence, and that the case is actually urgent enough, to justify such an extreme action. Of course, this is exactly the intended behavior, i.e. those are exactly the conditions under which the AI should really do something drastic.

As a more subtle case, consider an agent which has been strongly instructed to turn off, but which believes that it could serve its users better by continuing to operate. A very similar analysis applies to this case—if the user decides to reward shutting down then the agent will shutdown, and that's that. The user might make reasonable exceptions for extreme cases where the AI has good information, but these exceptions will be informed by what the AI actually knows, and this looks like the intended behavior.

Although I believe that this approach achieves the intended *effect* of corrigible reasoning, I share the feeling that there is some unanswered question about "responsible self-doubt." I feel like there is some sense in which a reasoner can recognize that their reasoning is potentially flawed and behave conservatively as a result; approval-direction behaves conservatively, but it doesn't shed much light on the actual reasoning process. It just defers the problem to the human overseer. I don't know whether we will learn anything more by thinking about this philosophical angle; I think that it is somewhat promising, but it's not a priority for me.

Accessibility

If we define the goodness of a possible future by how much I would like it "upon reflection," then we must mean something extreme by "reflection." What do we need to do before we can think about and carefully evaluate a whole future, which may span billions of years, encompassing worlds as alien to us as we would be to the first farmers? This kind of "reflection" probably requires learning more than humanity yet knows, becoming radically different and wiser people, and resolving all outstanding questions about what kind of society we want to live in or what we value. In short, it is extreme.

By contrast, evaluating actions seems to be a much more modest goal. In order to evaluate an AI system's action "well enough" to guide its behavior, I need to know everything that AI system knows—and nothing more.

To illustrate, suppose that an AI is choosing between two boxes, A and B. One box contains a diamond, and we could learn this fact if we reflected long enough. But in order to incentivize an approval-directed AI to pick the correct box, we just need to figure out which box is more likely to contain the diamond, *given what the AI knows and how long it has to think*. If we think at least that long, using at least that much evidence, then "predicting what we will think" is just as good as "predicting which box actually contains the diamond."

So it seems like defining actions requires knowing only what the AI knows. This isn't a technical point—I think it has a massive effect on the feasibility of defining V.

A simple consequence is that we don't have to do any outlandish reflection, eliminating many possible failure modes.

But more important is that we can now *actually carry out this process of reflection in the real world*, by using AI systems to help us figure out what actions we approve of. This in turn allows us to use supervised learning, which immediately takes our control schemes from "very speculative" to "something we could literally build today." Of course this introduces further questions, about whether this kind of actually-implementable-reflection is good enough—can we actually use AI assistants to understand everything that our AI systems "understand"? I think that these are key questions for moving forward with AI control, and I discuss them in the penultimate section.

Concerns

It seems useful to split concerns with this proposal up into two parts. Define an optimizer as a box which takes as input some kind of description of a function f and finds an output x with a high value of $f(x)$.

- Even if we had a good optimizer, we would still need to **define V**.
- **Building an optimizer** may be nearly as hard as the whole AI control problem.

In both cases, there would be a precisely analogous problem for any attempt to build a goal-directed agent.

The concern isn't that these problems are *worse* for approval-directed agents. It is that these problems are equally serious for both approaches, and that confronting them is the real substance of the AI control problem. The approval-directed approach simply obfuscates these problems by pushing them into the internal organization of the agent, or into the definition of V.

My own take is that these other problems simply don't seem analogous to the problems faced by a rational agent; I intend to do very different research on these problems than I would do if I wanted to better understand rational agency.

Defining approval

In order to define *V,* I need to define "the user" and "upon reflection."

My preferred approach is to use supervised learning. This addresses both questions as well as some more practical concerns. But it's a huge change that raises a host of new issues, and so it seems good to discuss approval-direction in this more abstract setting, where it is more comparable to existing work on AI safety.

"The user." Understanding how to define "the user" seems to amount to understanding how learning works, especially unsupervised learning or language learning. I think that this is a very natural theoretical question in AI, which probably deserves some attention even setting aside concerns with AI control.

I discuss the problem a bit here. This is closely related to the problems described in MIRI's technical agenda as "multi-level world models," "operator modeling," "ontology identification," and "ambiguity identification" (though I would approach and frame the problem differently).

I don't see how this is at all related to the difficulties posed by rational agency; I think that it is a different issue that will need to be resolved on either approach.

"Reflection." Defining "reflection" seems to be a tricky philosophical problem. How do we start from a computational system and infer "what it really wants," or "what it would do if it didn't make mistakes," or anything at all like that?

Again, this seems to be pretty orthogonal to concerns about rational agency, and will need to be solved or dodged in any case. There is a superficial connection, where we might extract human values by understanding a human as a rational agent. But if we want to carry out that research program, we seem to need a better understanding of human limitations, not of rational agency itself. Moreover, this approach doesn't seem nearly as promising to me as one based on defining an explicit process of reflection based directly on human behavior.

We might be able to find some nice approach to this problem based on a better understanding of rational agency, but it's definitely not the most natural place to look.

On value inference. We might try to dodge both of these problems by looking at human behavior (or even the modern world) and trying to infer what values it is optimized for.

In order to use this approach for a goal-directed agent, we need to do that extrapolation extremely carefully, and we would need to be able to infer "idealized" preferences that are independent of our current limitations. But rather than extracting preferences over possible worlds, we could extract preferences over behaviors for our AI systems. These preferences might be suitable for an approval-directed approach even if the idealization was much more limited and the judgments were much less robust. So I think that value inference is not a significant consideration in favor of a goal-directed approach, if anything it seems much easier to use as part of an approval-directed approach or narrow value learning approach.

Internal optimization

In order to actually build this kind of approval-directed agent, we would need to write the code that optimized the approval function.

It is clearly very difficult to write this code—indeed, this includes the entire AI problem. But for now, we want to know: is this problem actually any easier than the full AI control problem? Will all of the traditional concerns with goal-directed agents emerge in the course of building this optimizer?

I don't yet understand why we would run into these particular issues. In this section I want to try to explain my thinking.

Two organizing principles.

One basic question is: if we want to build a system that robustly picks actions maximizing V, how do we analyze the behavior of its components? What do we prove about this system, and about how the behavior of the pieces relate to the behavior of the whole?

There are two very natural candidates:

- Internal decisions should be chosen to optimize the value $V(a)$, for the action a ultimately output by the system. (Each internal decision has a causal effect on the action a, so this isn't an especially subtle goal.) We can enlarge the definition of a to include everything the system does that might plausibly have an external effect, or even to include every internal decision.
- For each internal decision i, define a similar function V_i which ranks each possible action a_i that could be taken. In the same way that V encodes reasoning about the consequences of each possible action, V_i encodes information about the consequences of each internal action.

I think that the second is more desirable to the extent that it can be implemented. Of course we can understand algorithms like backpropagation as encoding simple heuristics about the relationship between V_i and V; in this way we can understand pretty much anything as fitting into the second framework. The concern would be that if these heuristics are too crude to capture the details of the functions V_i, then we may end up implicitly using the first standard.

But even the first standard doesn't really seem problematic to me, so let's talk about it for the rest of this section. I see two classes of concerns with using this as an organizing principle for the internal behavior of your system.

Formulating this standard requires understanding rational agency.

This internal organization is basically the same as for a goal-directed agent. So aren't we back to square one, and now forced to improve our understanding of rational agency?

I don't think so. Even if we use a very bad formulation of goal-directed behavior, it doesn't have obvious bad consequences. As discussed in the "advantages" section, the system is equipped to correct such errors over time, and by construction they can't easily lead to bad actions (since such actions will be scored badly).

In order to claim that a bad internal notion of goal-directed behavior would lead to bad behavior, you have to claim that the internal organization of the agent, over the course of a single action, will generate situations where e.g. a decision-theoretic error would lead to irreversible trouble. For the most part this just doesn't seem plausible.

(A candidate counterexample is extortion—perhaps someone can threaten e.g. the memory-management unit of our AI into forgetting a critical fact, by instantiating a bunch of copies of that memory-management unit in contexts where the forgetting will be rewarded. I don't take extortion concerns too seriously for the overall behavior of the agent, and I take the internal version significantly less seriously. I do think that it's something to think about, especially as we deal with the immediate problems and can afford to be increasingly paranoid. But if this is the strongest justification that can be given for understanding goal-directed behavior then I am unconvinced. This is related to the next section.)

Alternatively, one could claim: current standards for goal-directed behavior are not only potentially dangerously wrong, they are also woefully incomplete and hence unusable as a formal criterion.

This second response seems uncompelling to me; basically I don't see where the research could end up such that we would have much clearer formal targets than we currently do. If we accept that any formalization of goal-directed behavior would be OK, then I expect that "what we can get" is likely to be the main constraint on our theorem statements. (This is related to the next section.)

Can you actually use this standard?

Even if we adopt such a standard, it's not clear that we can actually use it to design an agent. For example, the best available algorithms may simply happen to be hard to analyze within our preferred framework.

This seems like a serious obstacle to applying formal methods, but I don't really see why a better understanding of rational agency would be helpful—the algorithms may be hard to analyze for *any* particular framework. Attacking this problem seems to require thinking about the particular algorithms that might be hard to analyze, and trying to analyze them, rather than thinking about what formal guarantees would be most desirable if they were attainable.

As best as I can tell, some people think that a clear understanding of rational agency will itself suggest natural algorithmic techniques that are both effective and inherently easy-to-analyze within the framework of rational agency. But I don't see much evidence for this view. It's certainly possible, and if everything else seemed doomed to failure I might take it more seriously, but for now I'd definitely classify it as "long shot."

Concerns with supervision

As mentioned in the previous section, my preferred approach to AI control involves leveraging supervised learning. In principle this is orthogonal from the decision to use approval-direction, but in practice it is closely related, since (1) approval-direction is amenable to this approach (and I list that as an advantage), (2) using supervised learning changes the nature of many of the concerns, (3) even if the overall system is unsupervised, internal components may be supervised and so be subject to some of these concerns.

Briefly summarizing these additional concerns:

- **Perverse instantiations.** If we build systems that care about reward signals, we introduce instrumental incentives to manipulate those signals. In the approval-directed setting these incentives operate only when the system's actions are *shown* to a human, rather than when they are actually implemented. But even in this case, these actions will be optimized so that merely viewing them will distort the human's judgment and potentially compromise the reward channel. Some researchers interested in AI safety consider this to be a deal-breaker.

- **Limited reflection**. When using supervised learning we have to actually implement the intended process of reflection (so that we can use the outputs as training data). This means that we have to use a more limited form of reflection, defined by relatively short (e.g. week-long) interactions between humans and existing AI systems, rather than by a (potentially quite long) process of hypothetical extrapolation. There is a big open question about whether this process of reflection can actually leave the human well-enough informed to evaluate possible decisions for the AI.

- **Simulations**. When using supervised learning, it may be hard to train a system to predict what it will "actually" observe, if that is different from what will be observed by the vast majority of

simulated copies of that system. It's hard to know how serious a problem this will be. Moreover, many actors may be highly motivated to influence the predictions of powerful systems if it can be done cheaply.

Conclusion

Some research in AI control seems tightly wedded to a particular picture of rational agency; I suggest that some of these research may not be necessary, and offer approval-direction as an example of an approach that avoids it. (I also expect that practical designs for rational agents could dodge these issues, for example by inferring the user's instrumental preferences.)

In this post I discussed some reasons for skepticism with the approval-directed approach; I agree that these arguments suggest that more work is needed, but as far as I can tell this work is orthogonal to the distinction between approval-directed and goal-directed agents—it will be needed on both approaches, and won't significantly benefit from theoretical research on rational agency per se.

At the same time, approval-directed agents are potentially compatible with supervised learning, which seems like a key feature for building practical systems. Concerns with actually using supervised learning seem like key safety issues for a broad class of practical approaches to AI, and so addressing and understanding those concerns seems like it should be a priority.