Challenges for safe AI from RL

Paul Christiano Follow
Apr 9, 2015 · 10 min read

In this post, I'll describe and discuss a few big problems for the proposal from my last post. My overall sense is that none of these issues are fundamentally deal-breakers. But there are enough possible issues that I probably wouldn't want to bank on the proposal, and I think that in the long run it would be much better to find actual security arguments rather than a non-exhaustive list of possible problems that don't seem fatal.

I think that the most troubling scenario is outlined in the last problem, and I'll probably discuss this troubling scenario in more detail in a future post.

Problems

Problem 1: Maximization

The most obvious problem is that our AI is incentivized to find answers which receive the highest possible score, rather than answers which are actually good. We hope that these are answers which we would judge to be good, given careful judgment and the available facts.

But we may receive answers which are malicious in one way or another. They may exploit our misconceptions or biases, leverage other factors in our decision-making (e.g. by establishing credible threats of retaliation for a low rating), or could sieze control of their own reward channel by a physical or computer security exploit.

These problems become more serious when we attempt to evaluate actions by seeing how they play out, or by waiting until the future when we may be better-informed in general. But they are issues even if we are merely evaluating arguments presented as text.

Bootstrapping

Our main defense is applying other AIs to help protect ourselves from exploits. While a radically smarter AI may be able to find a "bad"

answer which convinces or coerces us into supplying a high rating, it will have a much harder time fooling a similarly intelligent peer.

We can use such bootstrapping in a two ways:

1. We can harden systems in advance based on advice from an AI, patching security loopholes, possible social exploits, or possible biases or lapses in judgment. This takes place in advance of making a query.

2. We can consult an AI to help evaluate a proposed action, by asking whether this particular proposal is in any way manipulative. We can then evaluate the claim of manipulativeness in the same way that we evaluate other claims, and if we conclude that an action was manipulative, we can adjust payoffs accordingly. This all takes place in step 3 of the process from the last post, though we may want to modify the process by allowing Hugh to pose individual questions to a predictor that is dumbed-down in one way or another.

Note that step 2 can take place before we actually look at a message or process it in any non-trivial way—we could have a sanitizing AI look at the message and try to detect and convince us of possible exploits, before we look at the message.

I see two major concerns with this approach.

First, it may be very difficult to harden an evaluation procedure against attempted manipulation. For example, it may be that even an AI's peers cannot adequately evaluate its arguments without being vulnerable to manipulation. In this case we could attempt to evaluate arguments with the assistance of AIs which are in some respects more powerful than the original arguer, but overall the prospects for argument look rather bleak in this world. I don't find this story very plausible or concerning, but the issue may not be settled until the empirical evidence is in.

Second, the AIs we consult may themselves manipulate us. To avoid this problem, we need these AIs to be "safer" than the AI whose argument we are evaluating. We can then defend ourselves against their manipulations using an even safer AI, until the process bottoms out with an AI which we can expose directly to human evaluators.

These assistant AIs can be safer by being less smart, by having less time to think, by being subject to more conservative evaluation by us,

by being restricted to make easier-to-evaluate arguments, by being forced to send simpler or shorter messages, or so on.

The key assumption is that notifying us of a possible exploit is "easier" than actually exploiting it, in at least one of these senses. That is, it can be done with less time to think or with less clever thoughts, it can be communicated in a less expressive medium or shorter message, or it can be correctly identified as a possible exploit by a more conservative evaluation process. Note that the notifier need not be able to work out the details of the exploit, they simply have to identify a possible vulnerability that shouldn't exist and might be exploitable by a sufficiently powerful attacker.

This seems very plausible to me. My main reason for optimism is that I haven't encountered any possible errors that weren't easier to point out than to exploit, in at least one of the above senses.

This process may sound unreasonably expensive, but it would only be applied rarely if ever. Each time we patch a vulnerability we can hopefully learn from our mistakes, and hopefully we don't have to repeat the process that many times. Moreover, it may be that most vulnerabilities are never even patched; the mere possibility of patching them when prompted could be sufficient to dissuade a manipulator from trying to exploit one.

Safety valves

A second defense is to ensure that manipulative answers are non-catastrophic, and to fix up problems once they are identified.

To some extent this is an automatic feature of our protocol, in which each learner wants to maximize their payoff but doesn't have greater ambitions to take over the world.

We could try to make this more robust by inserting safety valves: giving AIs every possible opportunity to receive payoff by making an attack in a non-destructive context. For example, we could periodically host arguments in which one side had a significant cognitive or structural advantage (more time to think, more information, a more expressive communication channel) but was saddled with defending an obviously false claim. If an agent were able to win one of these arguments, it would highlight a flaw which could subsequently be corrected, while (typically) causing very little collateral damage.

If we really have good episodic RL agents, they won't forego such opportunities in order to receive a higher payoff in future episodes: if they did that too many times, then we'd have a sequence of episodes in which their total payoff can be significantly increased by a simple modification, contradicting their goodness.

Problem 2: Uneven abilities

Weak understanding of humans

An AI may be very good at some tasks before it is able to reason effectively about humans. Such an AI may be able to devise very effective strategies for expanding their own influence, without being able to communicate effectively with humans or to predict what humans would or would not approve of if they were fully informed.

This seems to be a possible problem, but there are a few safeguards that leave me relatively unconcerned:

1. Reasoning about humans is an important part of interacting with the world, and one that humans are very keen to see AI excel at. So there is a good chance that it will be developed alongside other capabilities.
2. Even given a relatively modest ability to reason about humans and a very great ability to reason about other things, it is still possible to express arguments and explain plans. Indeed, if the only generalization you can make about humans is "they are magically able to figure out whether something is true by thinking about it," together with a very crude model of their preferences, then the proposals from my previous post will probably be effective. So the only really problematic case seems to be the one where the ability to reason about humans is *very* minimal. Given the current trajectory of AI, this doesn't seem like a very likely capability state. I discuss this issue briefly here.
3. An AI does not need to have a very detailed model of human preferences at all to pursue them relatively effectively. See the next section.

Weak arguers

Our proposal relies heavily on the ability of the AI to justify its behavior, as well as actually behaving well. It's still not clear whether an AI would necessarily have this ability, even given an adequate ability to predict human responses to possible arguments. This is

closely related to the concerns in problem 4. It's fine if this is much harder for them, as long as they can do it when called upon. And there is a great deal of latitude in the format of those arguments. But still, it's going to be hard to address this concern except with empirical evidence.

Other

Of course there may be other respects in which uneven abilities cause a problem. For example, an AI may just be really good at making explosives, and not so good at building anything useful. I think this is a very different issue which shouldn't be addressed at the same time, or with the same methods, as other concerns with AI risk.

Problem 3: Human values as a handicap

There are several respects in which an AI pursuing human values might operate at a fundamental handicap:

1. The AI may not know what outcomes are good or bad according to human values, and acquiring that information from humans may be a bottleneck.
2. Even if human values are understood well, it may be harder to figure out how to promote human values than how to acquire resources.
3. Acquiring resources most effectively may fundamentally involve compromising human values.

Issue #3 is mostly orthogonal to AI. That is, we may live in a world in which the most effective way to gain power is to commit acts of war, to pursue risky strategies which endanger life on Earth, or to do other terrible things. In that case we'll probably have to coordinate to prevent terrible things from happening, whether or not we build AI. AI doesn't seem to significantly change the difficulty of that coordination problem, except insofar as it generally improves humans' abilities to realize their goals (including both the goal of coordination and the goal of wanton destruction).

It seems that we can sidestep issues #1 and #2 if we are willing to accept a very slight loss. Namely, we can imagine an AI which pursues a simple proxy for human values: acquire resources, don't do anything really objectionable, keep the user safe (or humanity, or

whatever other group you represent), and eventually consult them about what to do with the resources.

It doesn't take much interaction with humans to determine that this is a reasonably good strategy, and to find a "good enough" model of really objectionable behavior (probably it doesn't take any interaction with humans at all, just reading things we've written would be plenty). And this strategy isn't much more complex than "acquire resources." I don't think an AI would actually follow a strategy like this—I think it would follow heuristics that are even simpler and even less objectionable—but this is a lower bound.

Formalizing this proxy precisely would be somewhat difficult, because of the complexity of the mandates "consult the user…" and "keep the user safe…" and even "acquire resources…" But it is an information-theoretically and computationally simple goal, and if an AI is able to reason about the world effectively it can presumably identify and pursue these kinds of simple subgoals. So whatever our AI does, it ought to receive feedback at least as good as if it implemented a strategy of this form—and if the evaluator is doing their job well, this means that the resulting outcomes are at least as good as if the AI implemented a strategy of this form.

Overall, I am not concerned that respecting human values will be a major handicap for an AI.

Problem 4: Inscrutability

In our protocol, most of the actual agency resides with the human users, in their counterfactual evaluations of actions. In order for this to work, the users need to actually be able to tell what actions further their values, if the reasoning is explained to them in sufficient detail (perhaps using the kinds of tricks described here and here, to "explain" things in a normally unattainable level of detail).

Some reasoning processes, like most search or inference processes in a fixed model, seem to produce comprehensible results. With sufficient time, we could inspect the results and verify that they are good explanations for the data or good plans to achieve the desired ends.

But other processes may produce completely inscrutable outputs. For example, an AI may learn a model for the world that has a great deal of latent structure. That latent structure need not correspond to any

human concepts. So while some intermediate steps of plans may be expressible in terms that a human can understand, some steps may only make sense in terms of completely alien concepts.

It may be possible to explain these alien concepts with enough time, or to convince a human that they relate to each other in certain ways, but this isn't guaranteed. Good reasoning may be fundamentally inscrutable.

Of course the process that produced the reasoning may be comprehensible anyway. We can be convinced that a latent variable has explanatory power, and that it is most likely to take on a particular value in a particular case, even without understanding anything about the meaning of that variable. In many cases, that would be enough to let us evaluate a plan. We seem most likely to run into trouble when the reasoning process really blurs the lines between prediction and planning; where even the internal thought process reveals no useful predictions beyond "This plan will be awesome."

The very worst case

In the very worst case:

- The most productive plan of action is incomprehensible.
- The process that produced that plan is equally incomprehensible.
- The safety of the plan is incomprehensible: the only way we can know it will be safe is to observe after the fact that no one died and got replaced by an adversarial superintelligence. Note that this only a problem for plans which may be *unusually* unsafe. A plan which merely killed everyone would be fine, because it wouldn't cause us to incorrectly rate the plan favorably in retrospect.
- The plan cannot be made safe with the assistance of any AI acting in a less-incomprehensible way. The only way to compete with incomprehensibility is with more incomprehensibility.

I don't know if this sort of thing is likely to happen; it seems unlikely to me, but I don't have any really compelling argument. If this is the nature of good plans, it seems to present a serious challenge for many approaches to AI safety.