A possible stance for alignment research

Paul Christiano Follow
Nov 30, 2015 · 9 min read

I think that AI alignment research should focus on building scalably aligned versions of contemporary systems—i.e. ML systems that are just as competent as unaligned systems, but which can are "trying to do the right thing."

By "scalable" I mean approaches that will continue to work just as well (and preferably better) as AI capabilities improve, as opposed to approaches that become increasingly unstable or unwieldy as AI improves.

In practice, that means we should (a) develop approaches to alignment for prosaic AI that appear *in principle* to be scalable, and (b) actually get them to work in practice. As easy versions of this goal are achieved, researchers can gradually work with weaker or more pessimistic assumptions, increasing the difficulty of the problem and the applicability of solutions.

In the rest of this post I'll say a bit about where I am coming from and explain my view in slightly more detail. I think the biggest uncertainty is how much to focus on contemporary AI systems.

AI control for now or then?

Roughly, we could identify two extremes on a spectrum:

1. **Short-term.** Work focused on the control problem posed by foreseeable techniques in the near future.
2. **Long-term.** Work focused on the control problem posed by unknown techniques in the further future.

Of course, many research directions might fall into both categories, and there is no bright line. But I find the breakdown useful, and I think that focusing on a different timescale suggests different priorities.

Considerations

Arguments for focusing on the short-term:

- It's easier to work with foreseeable techniques. We can focus our efforts more tightly, do empirical research, and think very concretely about possible designs rather than needing to work abstractly.
- Our contributions to AI control are much more relevant if the problem becomes relevant soon. If AI control is not relevant for many decades, then any work we do today will likely be subsumed by improved future understanding; in the "not relevant for a long time" scenario, there will be many years of work on AI control, that work will be better-informed by progress in AI, and the community working on the problem is likely to be much larger.
- Foreseeable techniques can serve as a concrete model for future techniques; even if this model is very likely to be wrong, it's not clear that we should expect it to be any more wrong than any particular set of assumptions about what future AI will look like.
- Work that engages with existing practice is much more likely to form a meaningful link between research in AI and AI control, and such a link seems desirable. On top of this effect, tractability is likely to provide a clearer model and inspiration for future work and to be generally good for the health of research on AI control.

Arguments for focusing on the long-term:

- AI probably won't be developed soon, and probably won't have a strong relationship with current techniques. Work predicated on these assumptions is most likely to be irrelevant, while we may be able to identify more general conceptual difficulties that are more "timeless," applicable to a wide range of possible architectures.
- If significant theoretical progress is needed in order to resolve the AI control problem, then that progress may have a substantial and hard-to-reduce "serial" component. For example, it is not clear that we could pursue meaningful theoretical research in this area without a basic understanding of probability theory and logic. By the same token, we might be

concerned that future research could end up blocking on similar developments that could be made today.

- Work that is only relevant in the farther future is much less crowded—many researchers do empirical research on existing AI systems, while rather few work on foundational philosophical issues. So to the extent that philosophical progress is necessary, it seems like a sorely neglected area.

Weighing up

On balance I think that the above considerations favor a short-term focus. Of course both forms of research are subject to diminishing returns and so both will be part of the optimal basket. But I suspect that the short-term focus should predominate, and is more useful on the margin.

Crowdedness vs. communication: Many more researchers do work related to practical AI systems than on foundational philosophical questions. However, only a vanishingly small fraction of this work is directly aimed at control—even projects that are nominally relevant to AI control are typically only partially relevant. I think that a researcher who is especially interested in scalable approaches to control will find no more than a handful of full-time equivalent researchers working on the most promising problems.

The existence of AI researchers in *closely related* practical fields seems to be a benefit of the short-term approach—not a cost. In the first place, these researchers may decide to contribute to the project. In the second place, these researchers are building intelligent systems, are most aware of the technological landscape and will be ultimately be responsible for integrating insight about control into practical designs; integration between control and capability has direct safety benefits beyond shared contributions.

Nearsightedness vs. improbability of AI soon. The AI control problem is not likely to be relevant soon, nor to be relevant to systems using (only) the particular techniques which are currently most promising. Work predicated on these assumptions is especially likely to be irrelevant if AI progress is slow, if the AI control problem is not relevant for a while, or if very different techniques become relevant in the interim.

But current work targeting future AI systems also makes significant additional assumptions about what they will look like and the nature

of the control problems that they will pose. At this point, those assumptions look at least as problematic as the assumptions of AI-control-problem-soon. And even conditioned on AI being developed far in the future, existing systems provide some evidence about the character of future AI systems.

Comparing these considerations requires some complex judgments about the trajectory of AI and the particular assumptions being made by future-focused control work. My take would be that near term work and far work are on comparable terms here.

In addition to this broader form of near-sightedness, we have a lot of detailed information about existing techniques, and the ability to do empirical research. So I think that on balance these two considerations point in favor of a short-term focus.

Serial progress vs. future crowdedness. I am pretty persuaded by the argument: "if AI control becomes an issue further in the future, more work will have been done by a larger community of better-informed researchers, and so our contributions matter less."

The countervailing argument is that, if these workers will be tied up in hard-to-parallelize research with inherent serial dependencies, then starting on that process earlier could add significant value—perhaps value that scales like the size of the future research community.

Setting aside the empirical question about whether there are such serial dependencies, even an infinitely strong view of "inherent serialness" merely negates the effect of future increases in the size of the research community working on AI control. It leaves in place the longer serial time available for research on the future problem, and the better information and tools available to the future community.

So it seems clear that work targeted for the near term will (all else equal) have more value, and the only question is whether it will have modestly more value or immensely more value. A strong view about serial progress is needed even to argue for "modestly more value." I don't think that a strong view about serial progress is warranted by the evidence, so I am even more relatively pessimistic about a long-term focus.

(One could also try to defend future research on the grounds that there has already been considerable safety work directed at current capabilities in particular—I don't think anyone believes this—or that

there was a long enough chain of clearly-necessary developments that doing safety work for the short term is a lost cause—I think that some people believe this, but I'm highly skeptical.)

Looking back

I've encountered the following rebuttal to this view: "if you had taken this perspective historically, you would have been thinking about AI control for expert systems, and today it would not do much good."

Thinking about this historical case has the opposite effect on my intuitions. Research on AI control for expert systems sounds pretty good to me, compared to a comparable head start on foundational questions in AI control. I'm not sure how to get at these intuitive disagreements. They could be disagreements about:

- The plausibility of GOFAI scaling far past human level using substantively similar techniques (so that practically-grounded AI control research would remain relevant). Are we talking 3%? 10%? 30%? I think 10% sounds quite pessimistic *ex ante*, but its a large enough probability to be a huge consideration.
- The relevance of research on "AI control for expert systems" to contemporary AI systems; how does this compare with the relevance of more abstract research that we could currently do? To me it looks like about a wash; both sets of assumptions seem poorly suited to contemporary systems, and I expect the benefits of concreteness to be comparable to the benefits of greater theoretical generality.
- The relevance of the 10 year head start on foundational questions. To me this looks pretty mild—it would not be hard to catch up by putting in some additional work now. The total impact seems negligible compared to the impact of 10 years of work in the 1970's, if it had become relevant in the 1980's.

But as a practical source of evidence about what to do, I don't know which of these things are the core disagreements and I suspect there are other factors at work.

What "AI control for now" might look like

Supposing that we want to focus on the AI control problems that may be posed by AI systems in the relatively near term, using foreseeable

techniques. What might such a focus look like?

When focusing on the near term, it is feasible to actually implement and test techniques for AI control in the context of existing AI systems. This seems like a good idea for a host of straightforward reasons, principally that experiment is an effective complement to theory and a prerequisite for meaningful integration with AI practice. It seems especially important in order to capture the hoped-for benefits of focusing on the near future.

Some solutions to the AI control problem work very well now but seem increasingly shaky as AI capabilities increase. For example, an approach that requires users to understand and provide reliable feedback on an agent's behavior after the fact may encounter trouble as AI improves and the implicit assumptions become more questionable.

This is why I remain interested in AI control even though it is currently a minor issue; we are really after scalable approaches to the AI control problem, that will continue to work just as well (or preferably better) as AI capabilities improve. I think this is an important concept that is worth fleshing out in more detail, but for now I'll use it informally.

Evaluating the scalability of an approach to AI control requires some amount of theoretical analysis. The availability of concrete implementations, together with experience applying those approaches at existing capability levels, seems to make that theoretical analysis much easier, and certainly makes it much more similar to theoretical discussion in the contemporary AI community.

This theoretical analysis can be made increasingly pessimistic along a number of axes: we can accommodate a wider range of possible futures, we can be more pessimistic about what will actually work and what might go wrong, we can minimize the amount of additional work that needs to be done in the future; or so on.

On this picture, the main distinguishing feature of AI control research as such is a focus on scalability. Over the long term, this is a natural goal to be absorbed by the broader AI research community, at which point there would be no need for this kind of control research as a separate endeavor.

I haven't searched very hard for alternative pictures of how AI control research could work, and realistically I don't think that actual events

will be well-described by such a simple picture. But I do think it is helpful to have the simple picture available, and I think it is a reasonable guide to prioritization for now.