Prosaic AI alignment

Paul Christiano Follow
Nov 19, 2016 · 10 min read

(*Related: a possible stance for AI control.*)

It's conceivable that we will build "prosaic" AGI, which doesn't reveal any fundamentally new ideas about the nature of intelligence or turn up any "unknown unknowns." I think we wouldn't know how to align such an AGI; moreover, in the process of building it, we wouldn't necessarily learn anything that would make the alignment problem more approachable. So I think that understanding this case is a natural priority for research on AI alignment.

In particular, I don't think it is reasonable to say "we'll know how to cross that bridge when we come to it," or "it's impossible to do meaningful work without knowing more about what powerful AI will look like." If you think that prosaic AGI is plausible, then we may already know what the bridge will look like when we get to it: if we can't do meaningful work now, then we have a problem.

1. Prosaic AGI

It now seems possible that we could build "prosaic" AGI, which can replicate human behavior but doesn't involve qualitatively new ideas about "how intelligence works:"

- It's plausible that a large neural network can replicate "fast" human cognition, and that by coupling it to simple computational mechanisms—short and long-term memory, attention, etc.—we could obtain a human-level computational architecture.
- It's plausible that a variant of RL can train this architecture to actually implement human-level cognition. This would likely involve some combination of ingredients like model-based RL, imitation learning, or hierarchical RL. There are a whole bunch of ideas currently on the table and being explored; if you can't imagine any of these ideas working out, then I feel that's a failure of imagination (unless you see something I don't).

We will certainly learn something by developing prosaic AGI. The very fact that there were no qualitatively new ideas is itself surprising. And beyond that, we'll get a few more bits of information about which particular approach works, fill in a whole bunch of extra details about how to design and train powerful models, and actually get some experimental data.

But none of these developments seem to fundamentally change the alignment problem, and existing approaches to AI alignment are not bottlenecked on this kind of information. Actually having the AI in front of us may let us work several times more efficiently, but it's not going to move us from "we have no idea how to proceed" to "now we get it."

2. Our current state

2a. The concern

If we build prosaic superhuman AGI, it seems most likely that it will be trained by reinforcement learning (extending other frameworks to superhuman performance would require new ideas). It's easy to imagine a prosaic RL system learning to play games with superhuman levels of competence and flexibility. But we don't have any shovel-ready approach to training an RL system to autonomously pursue our values.

To illustrate how this can go wrong, imagine using RL to implement a decentralized autonomous organization (DAO) which maximizes its profit. If we had very powerful RL systems, such a DAO might be able to outcompete human organizations at a wide range of tasks—producing and selling cheaper widgets, but also influencing government policy, extorting/manipulating other actors, and so on.

The shareholders of such a DAO may be able to capture the value it creates as long as they are able to retain effective control over its computing hardware / reward signal. Similarly, as long as such DAOs are weak enough to be effectively governed by existing laws and institutions, they are likely to benefit humanity even if they reinvest all of their profits.

But as AI improves, these DAOs would become much more powerful than their human owners or law enforcement. And we have no ready way to use a prosaic AGI to actually represent the shareholder's interests, or to govern a world dominated by superhuman DAOs. In general, we have no way to use RL to actually interpret and implement human wishes, rather than to optimize some concrete and easily-calculated reward signal.

I feel pessimistic about human prospects in such a world.

2b. Behaving cautiously

We could respond by not letting powerful RL systems act autonomously, or handicapping them enough that we can maintain effective control.

This leads us to a potentially precarious situation: everyone agrees to deploy handicapped systems over which they can maintain meaningful control. But any actor can gain an economic advantage by skimping on such an agreement, and some people would prefer a world dominated by RL agents to one dominated by humans. So there are incentives for defection; if RL systems are very powerful, then these incentives may be large, and even a small number of defectors may be able to rapidly overtake the honest majority which uses handicapped AI systems.

This makes AI a "destructive technology" with similar characteristics to e.g. nuclear weapons, a situation I described in my last post. Over the long run I think we will need to reliably cope with this kind of situation, but I don't think we are there yet. I think we could *probably* handle this situation, but there would definitely be a significant risk of trouble.

The situation is especially risky if AI progress is surprisingly rapid, if the alignment problem proves to be surprisingly difficult, if the political situation is tense or dysfunctional, if other things are going wrong at the same time, if AI development is fragmented, if there is a large "hardware overhang," and so on.

I think that there are relatively few plausible ways that humanity could permanently and irreversibly disfigure its legacy. So I am extremely unhappy with "a significant risk of trouble."

2c. The current state of AI alignment

We know many *approaches* to alignment, it's just that none of these are at the stage of something you could actually implement ("shovel-ready")—instead they are at the stage of research projects with an unpredictable and potentially long timetable.

For concreteness, consider two intuitively appealing approaches to AI alignment:

- **IRL:** AI systems could infer human preferences from human behavior, and then try to satisfy those preferences.
- **Natural language:** AI systems could have an understanding of natural language, and then execute instructions described in natural language.

Neither of these approaches is shovel ready, in the sense that we have no idea how to actually write code that implements either of them—you would need to have some good ideas before you even knew what experiments to run.

We might hope that this situation will change automatically as we build more sophisticated AI systems. But I don't think that's necessarily the case. "Prosaic AGI" is at the point where we can actually write down some code and say "maybe this would do superhuman RL, if you ran it with enough computing power and you fiddled with the knobs a whole bunch." But these alignment proposals are nowhere near that point, and I don't see any "known unknowns" that would let us quickly close the gap. (By construction, prosaic AGI doesn't involve unknown unknowns.)

So if we found ourselves with prosaic AGI tomorrow, we'd be in the situation described in the last section, for as long as it took us to complete one of these research agendas (or to develop and then execute a new one). Like I said, I think this would *probably* be OK, but it opens up an unreasonably high chance of really bad outcomes.

3. Priorities

I think that prosaic AGI should probably be the largest focus of current research on alignment. In this section I'll argue for that claim.

3a. Easy to start now

Prosaic AI alignment is especially interesting because the problem is nearly as tractable today as it would be if prosaic AGI were actually available.

Existing alignment proposals have only weak dependencies on most of the details we would learn while building prosaic AGI (e.g. model architectures, optimization strategies, variance reduction tricks, auxiliary objectives…). As a result, ignorance about those details isn't a huge problem for alignment work. We may eventually reach the point where those details are critically important, but we aren't there yet.

For now, finding *any* plausible approach to alignment, that works for *any* setting of unknown details, would be a big accomplishment. With such an approach in hand we could start to ask how sensitive it is to the unknown details, but it seems premature to be pessimistic before even taking that first step.

Note that even in the extreme case where our approach to AI alignment would be completely different for different values of some unknown details, the speedup from knowing them in advance is at most 1/(probability of most likely possibility). The most plausibly critical details are large-scale architectural decisions, for which there is a much smaller space of possibilities.

3b. Importance

If we do develop prosaic AGI without learning a lot more about AI alignment, then I think it would be bad news (see section **2**). Addressing alignment earlier, or having a clear understanding of why it intractable, would make the situation a lot better.

I think the main way that an understanding of alignment could *fail* to be valuable is if it turns out that alignment is very easy. But in that case, we should also be able quickly to solve it now (or at least have some *candidate* solution), and then we can move on to other things. So I don't think "alignment is very easy" is a possibility that should keep us up at night.

Alignment for prosaic AGI in particular will be less important if we don't actually develop prosaic AGI, but I think that this is a very big problem:

First, I think there is a reasonable chance (>10%) that we will build prosaic AGI. At this point there don't seem to be convincing arguments against the possibility, and one of the lessons of the last 30 years is that learning algorithms and lots of computation/data can do surprisingly well compared to approaches that require understanding "how to think."

Indeed, I think that if you had forced someone in 1990 to write down a concrete way that an AGI might work, they could easily have put

10–20% of their mass on the same cluster of possibilities that I'm currently calling "prosaic AGI." And if you'd ask them to guess what prosaic AGI would look like, I think that they could have given more like 20–40%.

Second, even if we don't develop prosaic AGI, I think it is very likely that there will be important similarities between alignment for prosaic AGI and alignment for whatever kind of AGI we actually build. For example, whatever AGI we actually build is likely to exploit many of the same techniques that a prosaic AGI would, and to the extent that those techniques pose challenges for alignment we will probably have to deal with them one way or another.

I think that working with a concrete model that we have available now is one of the best ways to make progress on alignment, even in cases where we *are* sure that there will be at least one qualitative change in how we think about AI.

Third, I think that research on alignment is significantly more important in cases where powerful AI is developed relatively soon. And in these cases, the probability of prosaic AGI seems to be *much* higher. If prosaic AGI is possible, then I think there is a significant chance of building broadly human level AGI over the next 10–20 years. I'd guess that hours of work on alignment are perhaps 10x more important if AI is developed in the next 15 years than if it is developed later, just based on simple heuristics based on diminishing marginal returns.

3c. Feasibility

Some researchers (especially at MIRI) believe that aligning prosaic AGI is probably infeasible—that the most likely approach to building an aligned AI is to understand intelligence in a much deeper way than we currently do, and that if we manage to build AGI before achieving such an understanding then we are in deep trouble.

I think that this shouldn't make us much less enthusiastic about prosaic AI alignment:

First, I don't think it's reasonable to have a confident position on this question. Claims of the form "problem X can't be solved" are *really hard to get right*, because you are fighting against the universal quantifier of all possible ways that someone could solve this problem. (This is very similar to the difficulty of saying "system X can't be compromised.") To the extent that there is any argument that aligning prosaic AGI is infeasible, that argument is nowhere near the level of rigor which would be compelling.

This implies on the one hand that it would be unwise to assign a high probability to the infeasibility of this problem. It implies on the other hand that even if the problem is infeasible, then we might expect to develop a substantially more complete understanding of why exactly it is so difficult.

Second, if this problem is actually infeasible, that is an extremely important fact with direct consequences for what we ought to do. It implies we will be unable to quickly play "catch up" on alignment after developing prosaic AGI, and so we would need to rely on coordination to prevent catastrophe. As a result:

- We should start preparing for such coordination immediately.
- It would be worthwhile for the AI community to substantially change its research direction in order to avoid catastrophe, even though this would involve large social costs.

I think we don't yet have very strong evidence for the intractability of this problem.

If we could *get* very strong evidence, I expect it would have a significant effect on changing researchers' priorities and on the research community's attitude towards AI development. Realistically, it's probably also a precondition for getting AI researchers to make a serious move towards an alternative approach to AI development, or to start talking seriously about the kind of coordination that would be needed to cope with hard-to-align AI.

Conclusion

I've claimed that prosaic AGI is conceivable, that it is a very appealing target for research on AI alignment, and that this gives us more reason to be enthusiastic for the overall tractability of alignment. For now, these arguments motivate me to focus on prosaic AGI.