Directions and desiderata for AI alignment

Paul Christiano Follow
Feb 6, 2017 · 16 min read

In the first half of this post, I'll discuss three research directions that I think are especially promising and relevant to AI alignment:

1. **Reliability and robustness.** Building ML systems which behave acceptably in the worst case rather than only on the training distribution.
2. **Oversight / reward learning.** Constructing objectives and training strategies which lead our policies to do what we intend.
3. **Deliberation and amplification.** Surpassing human performance without simultaneously abandoning human preferences.

I think that we have several angles of attack on each of these problems, and that solutions would significantly improve our ability to align AI. My current feeling is that these areas cover much of the key work that needs to be done.

In the second half of the post, I'll discuss three desiderata that I think should guide research on alignment:

1. **Secure**. Our solutions should work acceptably even when the environment itself is under the influence of an adversary.
2. **Competitive**. Our solutions should impose minimal overhead, performance penalties, or restrictions compared to malign AI.
3. **Scalable.** Our solutions should continue to work well even when the underlying learning systems improve significantly.

I think that taking these requirements seriously leads us to substantially narrow our focus.

It may turn out that these desiderata are impossible to meet, but if so I think that the first order of business should be understanding clearly *why* they are impossible. This would let us better target our work on alignment and better prepare for a future where we won't have a completely satisfying solution to alignment.

(The ideas in this post are not novel. My claimed contribution is merely collecting these things together. I will link to my own writing on each topic in large part because that's what I know.)

I. Research directions

1. Reliability and robustness

Traditional ML algorithms optimize a model or policy to perform well on the training distribution. These models can behave arbitrarily badly when we move away from the training distribution. Similarly, they can behave arbitrarily badly on a small part of the training distribution.

I think this is bad news:

- Deploying ML systems will critically change their environment, in a way that is hard or impossible to simulate at training time. (The "treacherous turn" is a special case of this phenomenon.)
- Deployed ML systems are interconnected and exposed to the same world. So if conditions change in a way that causes one of them to fail, *many* systems may fail simultaneously.
- If ML systems are extremely powerful, or if they play a critical role in society, then a widespread failure may have catastrophic consequences.

I'm aware of three basic approaches to reliability that seem to me like they could plausibly scale and be competitive:

(*ETA: this list is superseded by the list in Techniques for Optimizing Worst-Case Performance. I removed consensus and added interpretability and verification. I don't discuss "learning the right model," which I still consider a long shot.*)

- **Adversarial training**. At training time, attempt to construct inputs that induce problematic behavior and train on those. Eventually, we hope there will be no catastrophe-inducing inputs left. We don't yet know what is possible to achieve. (Szegedy 2014, Goodfellow 2015)

- **Ensembling and consensus**. We often have confidence that there exists *some* models which will generalize appropriately. If we can verify that many models agree about an answer, we can be confident that the consensus is correct. If we use this technique, we will often need to abstain on unfamiliar inputs, and in order to remain competitive we will probably need to represent the ensemble implicitly. (Khani 2016)

- **Learning the right model**. If we understood enough about the structure of our model (for example if it reflected the structure of the underlying data-generating process), we might be confident that it will generalize correctly. Very few researchers are aiming for a secure / competitive / scalable solution along these lines, and finding one seems almost (but not completely) hopeless to me. This is MIRI's approach.

Usual caveats apply: these approaches may need to be used in combination; we are likely to uncover completely different approaches in the future; and I'm probably overlooking important existing approaches.

I think this problem is pretty well-understood and well-recognized, but it looks really hard. ML researchers mostly focus on improving performance rather than robustness, and so I think that this area remains neglected despite the problem being well-recognized.

(Previous posts on this blog: *red teams, learning with catastrophes, thoughts on training highly reliable models*)

2. Oversight / reward learning

ML systems are typically trained by optimizing some objective over the training distribution. For this to yield "good" behavior, the objective needs to sufficiently close to what we really want.

I think this is also bad news:

- Some tasks are very "easy" to frame as optimization problems. For example, we can already write an objective to train an RL agent to operate a profit-maximizing autonomous corporation (though for now we can only train very weak agents).

- Many tasks that humans care about, such as maintaining law and order or helping us better understand our values, are extremely hard to convert into precise objectives: they are inherently poorly-defined or involve very long timescales, and simple proxies can be "gamed" by a sophisticated agent.

- As a result, many tasks that humans care about may not get done well; we may find ourselves in an increasingly sophisticated and complex world driven by completely alien values.

So far, the most promising angle of attack is to optimize extremely complex objectives, presumably by learning them.

I'm aware of two basic approaches to reward learning that seem like they could plausibly scale:

- **Inverse reinforcement learning**. We can observe human behavior in a domain and try to infer what the human is "trying to do," converting it into an objective that can be used to train our systems. (Russell 1998, Ng 2000, Hadfield-Menell 2016)
- **Learning from human feedback**. We can pose queries to humans to figure out which behaviors or outcomes they prefer, and then optimize our systems accordingly. (Isbell 2001, Thomaz 2006, Pilarski 2011, Knox 2012)

These solutions seem much closer to working than those listed in the previous section on reliability and robustness. But they still face many challenges, and are not yet competitive, scalable, *or* secure:

- IRL requires a prior over preferences and a model of how human behavior relates to human preferences. Current implementations either only work in severely restricted environments, or use simple models of human rationality which cause the learner to attempt to very precisely imitate the human's behavior (which might be challenging or impossible).
- For similar reasons, existing IRL implementations are not able to learn from other data like human utterances or off-policy behavior, even though these constitute the largest and richest source of data about human preferences.
- Human feedback requires accurately eliciting human preferences, which introduces many complications. (I discuss a few easy problems here.)
- Human feedback is expensive and so we will need to be able to learn from a relatively small amount of labeled data.

Demonstrations are also expensive and so may end up being a bottleneck for approaches based on IRL though it's not as clear.

- Both imitation learning and human feedback may fail when evaluating a behavior requires understanding where the behavior came from. For example, if you ask a human to evaluate a painting they may not be able to easily check whether it is derivative, even if over the long run they would prefer their AI to paint novel paintings.

(I've described these approaches in the context of "human" behavior, but the expert providing feedback/demonstrations might themselves be a human augmented with AI assistance, and eventually may simply be an AI system that is aligned with human interests.)

This problem has not received much attention in the past, but it seems to be rapidly growing in popularity, which is great. I'm currently working on a project in this area.

(*Previous posts on this blog: the reward engineering problem, ambitious vs. narrow value learning, against mimicry, thoughts on reward engineering*.)

3. Deliberation and amplification

Machine learning is usually applied to tasks where feedback is readily available. The research problem in the previous section aims to obtain quick feedback in general by using human judgments as the "gold standard." But this approach breaks down if we want to exceed human performance.

For example, it is easy to see how we could use machine learning to train ML systems to make human-level judgments about urban planning, by training them to produce plans that sound good to humans. But if we want to train an ML system to make superhuman judgments about how to lay out a city, it's completely unclear how we could do it—without spending billions of dollars trying out the system's ideas and telling it which ones work.

This is a problem for the same reasons discussed in the preceding section. If our society is driven by systems superhumanly optimizing short-term proxies for what we care about—such as how much they impress humans, or how much money they make—then we are liable to head off in a direction which does not reflect our values or leave us in meaningful control of the situation.

If we lowered our ambitions and decide that superhuman performance is inherently unsafe, we would be leaving huge amounts of value on the table. Moreover, this would be an unstable situation: it could last only as long as everyone with access to AI coordinated to pull their punches and handicap their AI systems.

I’m aware of two approaches to this problem that seem like they could scale:

- **IRL [hard mode]**. In principle we can use IRL to recover a representation of human preferences, and then apply superhuman intelligence to satisfy those preferences much better than a human could. However, this is a much more ambitious and challenging form of IRL than is usually discussed, which remains quite challenging even when you set aside all of the usual algorithmic and statistical difficulties. (Jacob Steinhardt and Owain Evans discuss this issue in a recent post.)
- **Iterated amplification**. A group of interacting humans can potentially be smarter than a single human, and a group of AI systems could be smarter than the original AI system. By using these groups as “experts” in place of individual humans, we could potentially train much smarter systems. The key questions are how to perform this composition in a way that causes the group to implement the same preferences as its members, and whether the cognitive benefits for groups are large enough to overcome the overhead of coordination. (I discuss this approach here and in follow-up work.)
- **IRL for cognition**. Rather than applying IRL to a humans’ actions, we could apply it to the cognitive actions taken by a human while they deliberate about a subject. We can then use those values to execute a longer deliberation process, asking “what would the human do if they had more time to think / more powerful cognitive tools?” I think this approach ends up being similar to a blend of the previous two.

It’s completely unclear how hard this problem is or how far we are from a solution. It is a much less common research topic than either of the preceding points.

In the short term, I think it might be easier to study analogs of this problem in the context of human behavior than to attempt to directly study it in the context of AI systems.

Ought is a non-profit aimed at addressing (roughly) this problem; I think it is reasonably likely to make significant progress.

(*Previous posts on this blog: capability amplification, reliability amplification, security amplification, meta-execution, the easy goal inference problem is still hard*)

II. Desiderata

I'm most interested in algorithms that are secure, competitive, and scalable, and I think that most research programs are very unlikely to deliver these desiderata (this is why the lists above are so short).

Since these desiderata are doing a lot of work in narrowing down the space of possible research directions, it seems worthwhile to be thoughtful and clear about them. It would be easy to gloss over any of them as obviously unobjectionable, but I would be more interested in people pushing back on the strong forms than implicitly accepting a milder form.

1. Secure

Many pieces of software work "well enough" most of the time; we often learn this not by a deep analysis but by just trying it and seeing what happens. "Works well enough" often breaks down when an adversary enters the prediction.

Whether or not that's a good way to build AI, I think it's a bad way to do alignment research right now.

Instead, we should try to come up with alignment solutions that work in the least convenient world, when nature itself is behaving adversarially. Accomplishing this requires argument and analysis, and cannot be exclusively or based on empirical observation.

AI systems obviously won't work well in the worst case (there is no such thing as a free lunch) but it's reasonable to hope that our AI systems will never respond to a bad input by actively *trying* to hurt us —at least as long as we remain in physical control of the computing hardware, and the training process, *etc*.

Why does security seem important?

- It's really hard to anticipate what is going to happen in the future. I think it's easy to peer into the mists and say "well, hard to know what's going to happen, but this solution might work out OK," and then to turn out to be too optimistic. It's harder to make this error when we hold ourselves to a higher standard, of actually giving an argument for why things work. I think that this is a general principle for doing useful research in advance of when it is needed—we should hold ourselves to standards that are unambiguous and clear even when the future is murky. This is a theme that will recur in the coming sections.

- We are used to technological progress proceeding slowly compared to timescales of human judgment and planning. It seems quite likely that powerful AI will be developed during or after a period of acceleration, challenging those assumptions and undermining a traditional iterative approach to development.
- The world really does contain adversaries. It's one thing to build insecure software when machines have power over modest amounts of money with significant human oversight, it's another thing altogether when they have primary responsibility for enforcing the law. I'm not even particularly worried about human attackers, I'm mostly worried about a future where all it takes to launch attacks is money (which can itself be earned by executing attacks). Moreover, if the underlying ML is insecure and ML plays a role in almost all software, we are going to have a hard time writing any secure software at all.

(*Previous posts: security and AI alignment*)

2. Competitive

It's easy to avoid building an unsafe AI system (for example: build a spreadsheet instead). The only question is how much you have to sacrifice to do it.

Ideally we'll be able to build benign AI systems that are just as efficient and capable as the best AI that we could build by any means. That means: we don't have to additional domain-specific engineering work to align our systems, benign AI doesn't require too much more data or computation, and our alignment techniques don't force us to use particular techniques or restrict our choices in other ways.

(More precisely, I would consider an alignment strategy a success if the additional costs are sublinear: if the fraction of resources that need to be spent on alignment research and run-time overhead *decreases* as the AI systems become more powerful, converging towards 0.)

Why is competitiveness important?

A. It’s easy to tell when a solution is plausibly competitive, but very hard to tell exactly how uncompetitive an uncompetitive solution will be. For example, if a purported alignment strategy requires an AI not to use technique or development strategy X, it’s easy to tell that this proposal isn’t competitive in general, but very hard to know exactly how uncompetitive it is.

As in the security case, it seems very easy to look into the fog of the future and say “well this seems like it will probably be OK” and then to turn out to be too optimistic. If we hold ourselves to the higher standard of competitiveness, it is much easier to stay honest.

Relatedly, we want alignment solutions that work across an extremely large range of techniques not just because we are uncertain about which techniques will be important, but because generalizing across all of the situations we can foresee is a good predictor of working for situations we can’t foresee.

B. You can’t unilaterally use uncompetitive alignment techniques; we would need global coordination to avoid trouble. If we *don’t* know how to build competitive benign AI, then users/designers of AI systems have to compromise efficiency in order to maintain reliable control over those systems. The most efficient systems will by default be built by whoever is willing to accept the largest risk of catastrophe (or perhaps by actors who consider unaligned AI a desirable outcome).

It may be possible to avert this kind of race to the bottom by effective coordination by e.g. enforcing regulations which mandate adequate investments in alignment or restrict what kinds of AI are deployed. Enforcing such controls domestically is already a huge headache. But internationally things are even worse: a country that handicapped its AI industry in order to proceed cautiously would face the risk of being overtaken by a less prudent competitor, and avoiding *that* race would require effective international coordination.

Ultimately society will be able and willing to pay *some* efficiency cost to reliably align AI with human interests. But the higher that cost, the harder the coordination problem that we will need to solve. I think the research community should be trying to make that coordination problem as easy as possible.

(*Previous posts: prosaic AI alignment, a possible stance for AI control, efficient and safely scalable*)

3. Scalable

Over time, we are acquiring more data, more powerful computers, richer model classes, better optimization algorithms, better exploration strategies, and so on. If we extrapolate these trends, we end up with very powerful models and policies.

Many approaches to alignment break down at some point in this extrapolation. For example, if we train an RL agent with a reward function which imperfectly approximates what we want, it is likely to fail once the agent becomes sufficiently sophisticated—unless the reward function itself becomes more sophisticated in parallel.

In contrast, let's say that a technique is "scalable" if it continues to work just as well even when the underlying learning becomes much more powerful. (See also: Eliezer's more colorful "omnipotence test.")

This is another extremely demanding requirement. It rules out many possible approaches to alignment. For example, it probably rules out any approach that involves hand-engineering reward functions. More subtly, I expect it will rule out any approach that requires hand-engineering an informative prior over human values (though some day we will hopefully find a scalable approach to IRL).

Why is scalability important?

- As in the previous sections, it's easy to be too optimistic about exactly when a non-scalable alignment scheme will break down. It's much easier to keep ourselves honest if we actually hold ourselves to producing scalable systems.
- If AI progress rapidly, and especially if AI research is substantially automated, then we may literally confront the situation where the capabilities of our AI systems are changing rapidly. It would be desirable to have alignment schemes that continued to work in this case.
- If we don't have scalable solutions then we require a continuing investment of research on alignment in order to "keep up" with improvements in the underlying learning. This risks compromising competitiveness, forcing AI developers to make a hard tradeoff between alignment and capabilities. This would be acceptable if the ongoing investments in alignment are modest

compared to the investments in capabilities. But as with the last point, that’s a very murky question about which it seems easy to be overly optimistic in advance. If we think the problem will be easy in the future when we have more computing, then we ought to be able to do it now. Or at the very least we ought to be able to explain how more computing will make it easy. If we make such an explanation sufficiently precise then it will itself become a scalable alignment proposal (though perhaps one that involves ongoing human effort).

(*Previous posts: scalable AI control*)

Aside: feasibility

One might reject these desiderata because they seem too demanding: it would be great if we had a secure, competitive and scalable approach to alignment, but that might not be possible.

I am interested in trying to satisfy these desiderata despite the fact that they are quite demanding, for two reasons:

- I think that it is very hard to say in advance what is possible or impossible. I don’t yet see any fundamental obstructions to achieving these goals, and until I see hard obstructions I think there is a significant probability that the problem will prove to be feasible (or “almost possible,” in the sense that we may need to weaken these goals only slightly).
- If there is some fundamental obstruction to achieving these goals, then it would be good to understand that obstruction in detail. Understanding it would help us understand the nature of the problem we face and would allow us to do better research on alignment (by focusing on the key aspects of the problem). And knowing that these problems are impossible, and understanding exactly how impossible they are, helps us prepare for the future, to build institutions and mechanisms that will be needed to cope with unavoidable limitations of our AI alignment strategies.

III. Conclusion

I think there is a lot of research to be done on AI alignment; we are limited by a lack of time and labor rather than by a lack of ideas about how to make progress.

Research relevant to alignment is already underway; researchers and funders interested in alignment can get a lot of mileage by supporting and fleshing out existing research programs in relevant directions. I don't think it is correct to assume that if anyone is working on a problem then it is going to get solved—even amongst things that aren't literally at the "no one else is doing it" level, there are varying degrees of neglect.

At the same time, the goals of alignment are sufficiently unusual that we shouldn't be surprised or concerned to find ourselves doing unusual research. I think that area #3 on deliberation and amplification is almost completely empty, and will probably remain pretty empty until we have clearer statements of the problem or convincing demonstrations of work in that area.

I think the distinguishing feature of research motivated by AI alignment should be an emphasis on secure, competitive, and scalable solutions. I think these are very demanding requirements that significantly narrow down the space of possible approaches and which are rarely explicitly considered in the current AI community.

It may turn out that these requirements are infeasible; if so, one key output of alignment research will be a better understanding of the key obstacles. This understanding can help guide less ambitious alignment research, and can help us prepare for a future in which we won't have a completely satisfying solution to AI alignment.

This post has mostly focused on research that would translate directly into concrete systems. I think there is also a need for theoretical research building better abstractions for reasoning about optimization, security, selection, consequentialism, and so on. It is plausible to me that we will produce acceptable systems with our current conceptual machinery, but if we want to convincingly *analyze* those systems then I think we will need significant conceptual progress (and better concepts may lead us to different approaches). I think that practical and theoretical research will be attractive to different researchers, and I don't have strong views about their relative value.