Scalable AI control

Paul Christiano Follow
Dec 5, 2015 · 13 min read

By AI control, I mean the problem of getting AI systems to do what we want them to do, to the best of their abilities.

More precisely, my goal is minimizing the gap between how well AI systems can contribute to *our* values, and how well they can pursue *other* values.

Why might such a gap exist?

Depending on how AI develops, we may be especially good at building AI systems that pursue objectives defined directly by their experiences—as in reinforcement learning—or which have a simple explicit representation. If human values don't fit into these frameworks, the best AI systems may optimize simple proxies for "what we really want."

If those proxies aren't good enough? Then we have a gap.

Scalability

Whether or not such a gap *could* come to exist, today it doesn't seem to. So: is it possible to do empirical work on AI control today?

I think so. My preferred approach is to focus on **scalable** approaches: those that will continue to work, and which will preferably work *better*, as our AI systems become more capable. We should not be satisfied with control techniques that will eventually break down, or which will require continuous innovation in order to keep pace with AI capabilities.

This gives us something to do today: try to devise practical and scalable solutions to the AI control problem.

Even if we don't think that this is a good objective for current AI control research, I think that the concept of scalable AI control is very useful (and is probably more broadly acceptable than substitutes like "the superintelligence control problem" or "omni-safety").

This definition is not completely precise, and I provide some important clarification in Section 2. I think that it in practice there is often a pretty clear line between scalable and unscalable proposals, which I think can serve both as a useful concept and a useful research direction.

In Section 3 I'll provide some examples of approaches which aren't scalable, and in Section 4 I'll discuss some alternative goals for AI control research.

2. Clarification

Scaling to handle what?

It's only really meaningful to talk about whether a technique can be scaled to handle *some particular kind of progress*. There is no unified spectrum of "capability" that AI is progressing along steadily.

Ideally we would make techniques that can scale to handle any kind of progress that might occur. In practice I am interested in techniques that can handle various simple, foreseeable kinds of progress. Once that bar can be met, we can consider a wider and wider range of possible trajectories.

What are simple, foreseeable forms of progress? Easy examples are faster computers, better optimization strategies, or richer model classes that are easier to optimize over. More generally, researchers are working on and making continuous progress on a wide range of concrete problems, and in most cases we can imagine progress continuing for a long time, without fundamentally changing the nature of the AI control problem.

We can broaden the space of possible trajectories by considering progress on a broader range of existing techniques, including techniques that are currently impractical. We can also consider concrete future capabilities that might be attained. Or we can try to design control techniques that will extend to completely unanticipated developments, based increasingly minimal assumptions about the nature of those developments.

But for now, I think that scaling to handle concrete, foreseeable developments is a hard enough problem.

Example: reinforcement and supervised learning

There is an especially nice way to think about scalability with respect to progress in reinforcement and supervised learning.

These techniques produce systems that optimize an objective defined by explicit feedback. We can easily imagine better systems that more effectively optimize the given objective. And we can ask: do our control techniques work as our systems get better and better at optimizing these objectives, or are they predicated on implicit assumptions about the limitations of our systems? In the limit, we can consider systems that literally choose the output maximizing the given reward signal.

I think that this view of scalability is distinctive to MIRI, and I think it is a great aspect of their methodology. They would use a slightly different version of the principle, in which a system might be optimized for any precisely-defined objective, but the underlying principle is quite similar.

My version essentially amounts to assuming that (1) reinforcement learning, broadly construed, will remain a dominant methodology in AI, and (2) there will be no progress in reward engineering.

I think (1) is plausible though unlikely. I think that (2) is implausible. Ignoring future progress in reward engineering is a methodological approach, intended to help us understand the problem of reward engineering rather than to make a prediction. This brings us to:

The intended path of progress

My guess is that the null AI control policy—do nothing—would in fact scale to the actual AI progress that actually occurs.

This is just a restatement of my optimistic view of the world: I expect that we will be good at building AI systems that do the things we want them to do, by the time that we are good at building AI systems that do anything. If that’s how things go, then we wouldn’t need any additional insight to handle AI control, because by assumption there is no problem.

But my goal is to do work now that decreases the probability and extent of trouble. And from that perspective, it is quite natural to consider alternative (more problematic) trajectories for progress, and to focus on work we could do today that would make those trajectories non-problematic.

This is useful as a hedge against possible bad outcomes—the reason to work on AI control now is the possibility that it will eventually be a serious problem. But it's not merely a hedge.

As an analogy, suppose that I want to devise techniques to make cars safer. I work at a car company, which is currently designing our 2018 model, and I'm thinking of safety features for that car. I wouldn't say: "it seems like no further work is needed; obviously the 208 model will be built to incorporate all reasonable safety precautions." The whole point of the exercise is to think of technologies might make the car safer. We are imagining future cars in the interests of better understanding the safety problem.

I want to stress that thinking about these unfavorable trajectories *isn't a prediction about what AI progress will look like*. It's a methodological strategy for finding research problems that *improve the probability that AI progress will be robustly beneficial for humanity*.

Hard values, easy values

My definition of AI control may be difficult to achieve if our values are fundamentally harder to realize than some other, simpler values. This is a problem for my statement of the AI control problem: "minimizing the gap between [how well AI systems can contribute to *our* values] and [how well AI systems can contribute to *other* values]."

My statement of the control problem is only really meaningful because there are instrumental subgoals that are shared (or are extremely similar) between many different values, which let us compare the efficacy with which agents pursue those different values. Performance on these very similar subgoals should be used as the performance metric when interpreting my definition of AI control problem.

In fact even if we only resolved the problem for the similar-subgoals case, it would be pretty good news for AI safety. Catastrophic scenarios are mostly caused by our AI systems failing to effectively pursue convergent instrumental subgoals on our behalf, and these subgoals are by definition shared by a broad range of values.

Scalable with how much work?

Very few algorithms can be *literally* applied without modification to a radically different future setting. Obviously our goal is to minimize the work that would be needed to adapt a given control approach to improved future techniques. For example, a technique that increased the difficulty of deploying future AI systems by 1%, or which required a constant amount of work to apply existing AI techniques to new problems, would seem great.

The real problem is when scaling an AI control technique relies on future people discovering as-yet-unknown insights, doing an unknown and potentially large amount of additional work, or doing an amount of additional work that is large relative to the total quantity of AI research.

3. Example of non-scalable control techniques

Example: reinforcement learning

Consider a particularly simple technique for AI control. Start with a reinforcement learner. Give the user a button that controls the reinforcement learner's reward. The user can ask the reinforcement learner to perform tasks, and can provide a reward whenever it succeeds. The learner will hopefully learn how to interpret and satisfy these requests.

This would be a fine approach to AI control at the moment (though in fact we can't build systems clever enough that it's worth doing). However, this approach will work less well as AI systems improve:

- If the reinforcement learner is more competent than the user in a particular domain, then the user may not be able to provide good evaluations of the learner's behavior. For example, if the learner is a world-class party planner, they will be forced to produce a party plan that *looks good to the user* rather than one that would actually *be good*.
- All the reinforcement learner really cares about is the reward signal, not the attitude of the user. So the learner is liable to manipulate, deceive, or threaten the user into receiving additional reward.

By contrast, an increasingly powerful reinforcement learner would be much more effective while optimizing for its own physical security (in order to ensure that it continues to receive a high reward). For example, if its income depended on party-planning, it would apply its full party-planning abilities towards throwing a profit-maximizing party.

So I would say this control technique is not scalable. If we want to use this technique, we will either need to accept degraded performance on the AI control problem, or (more likely) continue to do additional work as AI capabilities improve in order to ensure that control "keeps up."

Example: imitation

Suppose that I want to train an AI to drive a car. A very simple procedure would be to copy human driving: have an expert drive a car, record the sensor readings and the expert's actions, and train a model which maps a sequence of sensor readings to predictions of the human's actions. We can then use this model to control a car by doing what the human is predicted to do.

Suppose for the sake of argument that this actually resulted in OK driving. (In fact it has a number of rather serious problems.)

No matter how good our learning system is, this procedure will never generate substantially superhuman driving. For example, even if a human is expecting the car in front of them to brake, it will still take them hundreds of milliseconds to actually respond to the brake lights. So a system trained to imitate human behavior will add a gratuitous delay before it responds to observed brake lights.

Using sophisticated learning systems, we could likely achieve better performance by specifying the *goal* of driving (don't crash, have a smooth ride, get to the destination quickly) and allowing the system to devise its own policies in order to achieve that goal.

So this approach to teaching a car to drive is also not scalable. As AI improves, systems trained using this technique will fall behind.

4. Alternatives

I think that building practical and scalable control systems is an especially good goal for organizing work on AI control. But there are many other possibilities (in addition to just playing it by ear). Here are a few alternatives that seem salient to me:

Pursue a long-term vision

The MIRI research agenda is built around a particular vision for how sophisticated AI might be aligned with human interests.

Once we have a vision in mind, we can search for concrete problems that would be needed to realize this vision. This is another perfectly good source of research projects that might help with control.

The main reason I'm less keen on this approach is that it puts a lot of weight on your long-term vision. Most researchers I know who object to MIRI's research agenda do so because they don't think that the long-term vision is especially plausible. If you depart at that stage, we don't really have any good "rules of the game" that can arbitrate the debate. Moreover, even if MIRI succeeds spectacularly at their research agenda, it won't really alleviate these concerns.

So it seems like if we want to take this route, a lot of the work is being done by the first step of the problem where we identify the long-term picture and the necessary ingredients. Given that that's where a lot of the actual work is getting done, I suspect it's also where most of the effort should go. But this cuts against "pursuing a particular long-term vision" as an organizing goal for research.

This isn't entirely fair, because pursuing a vision also contributes to testing the feasibility of that vision. I am more sympathetic to the kind of "pursuit" that also constitutes "testing," for this reason.

The steering problem

I previously suggested that researchers in AI control try to answer the question: "Using black-box access to human-level cognitive abilities, can we write a program that is as useful as a well-motivated human with those abilities?"

I've found this perspective to be very useful in my own thinking, and I continue to recommend it as a question to think about.

But it suffers from (1) a lack of precision, and (2) a reliance on well-defined black-box capabilities that may not match the actual capabilities we develop.

By working with the capabilities available today, we can largely sidestep these issues. Working with existing capabilities gives us an extremely precise and internally detailed model system.

When we think about scalability, and in particular about the kinds of progress that we should be able to cope with, we start to run into some of the same difficulties that afflict the steering problem. But these difficulties look much less precise, and much more closely wedded to actual AI progress in a way that makes it easier to agree about what kinds of extrapolation are reasonable and which are not.

Model problems

Another approach would be to construct toy domains in which the control problem, or some problem which we would judge to be analogous, is already difficult.

Hard cases: For example, the reinforcement learning and imitation solutions discussed in Section 2 don't really *perfectly* solve even the existing control problem. So we could focus on having very good solutions to the control problem in challenging domains, where human performance is worse than AI performance and where humans cannot easily evaluate the quality of a given performance.

I think that this is a basically reasonable direction for research, and I doubt it would look too different from thinking about scalable AI control. It forces us to consider a very narrow range of regimes and to confront only a small range of possible problems, which I think is something of a disadvantage. It also forces us to really play for "small stakes," since the failures of control in existing environments don't seem like a huge deal (and can largely be resolved by ad hoc measures). I think this is another disadvantage, but it might be partly alleviated by emphasizing the fact that these problems are expected to get worse and are worth attacking in advance.

This approach has the advantage that it's not totally dependent on any argument about "scalability," or really any complex argument at all. It is able to focus on concrete problems that exist today, which are basically guaranteed to be real problems. That said, the argument for these problems being important or especially interesting probably *does* rest on some kind of argument about the future.

Overall, I think that focusing on hard cases is reasonable, and is a useful complement even if we want to focus on implementing scalable solutions, as long as we can trace failures of scalability back to failures of existing systems in *some* domain.

Subproblems. A different approach would be to identify problems that we think are likely to be solved as part of a successful AI control approach. Those problems might not be resolved today, even if the AI control problem is. For example, we might think that value learning will necessarily play a much larger role in future systems than it plays today. But even today we aren't very good at value learning, and so this gives us a concrete problem to work on.

This seems basically equivalent to the "pursuing a long-term vision" point above, and has mostly the same advantages and disadvantages.

Analogies. A final approach is to consider problems which look like they are usefully analogous to the control problem, but which are currently significantly harder. This might give us a concrete model that exposes more of the difficulties of the control problem, and it might involve fewer assumptions than picking a long-term vision.

A simple example would be a game played by two AI systems. One is a reinforcement learner which has some hard-to-communicate information about what it wants done. The other is an AI assistant whose design is completely flexible, and which has significantly more computational power than the RL agent. (Alternatively or in addition, it may have some other resources that the RL agent lacks, like extra actuators.) Our goal is to specify the assistant, and provide a strategy for the RL agent, that allows the RL agent to achieve its goals nearly as effectively as if it had all of the resources of the assistant.

Because we can make the AI assistant much more powerful than the RL agent, this can allow us to capture some anticipated difficulties of AI control before we can actually build AI systems that are much more powerful than their human users.

I think that building this kind of analogy might be very useful for AI control, and it seems quite worthwhile. I think it is a substantially different approach than trying to work on scalable AI control, and it might turn out to be more promising.

However, there are many additional difficulties with setting up this kind of analogy, and I think there is a good chance that those difficulties will prove fatal. I think the biggest problems are that:

- There will be solutions that work in the analogy that won't work for the real control problem. In the example above, an RL agent might pass its reward signals on to the assistant, who could use standard RL algorithms to pursue them. Or there may be difficulties that depend on the absolute capability of the assistant rather than on the difference between its capabilities and those of the human.
- There will be many difficulties in the analogy that aren't difficulties in real life. In the example above, it might be quite hard to build the RL system that is supposed to be a model of humans, but this isn't really part of the AI control problem. Or an accurate model of the problem may require building systems that are actually embedded in the environment, and building such an environment may be a massive engineering challenge orthogonal to control.

Conclusion

I think scalability is a useful concept when reasoning about AI control. I think that designing practical but scalable AI control techniques is also a promising goal for research on AI control.

This post clarified the term, provided some examples, and discussed some alternative goals.