Technical and social approaches to AI safety

Paul Christiano Follow
Apr 12, 2015 · 7 min read

I often divide solutions to the AI control problem into two parts: technical and social. I think any solution requires progress on both fronts, but strength on one front can compensate for a weakness on the other.

This view suggests that most of the value comes from having a "good" technical solution (quantification below), rather than a "perfect" solution or a "not terrible" solution. It also suggests that there is value from progress on both technical and social solutions, rather than justifying a focus on one or the other.

Technical solutions

If we can build AI systems that are both efficient and safe, then we have addressed the most serious concerns about AI risk, regardless of how well or poorly society coordinates. If efficient AI is unsafe, then coordination may be needed to prevent people from using it.

1. **Efficient**: An efficient system is as cheap and effective, for the tasks to which it is applied, as any other system we can build using available technology. As AI becomes cheaper relative to human labor, the efficiency penalty for human involvement becomes larger. In the long run, efficient systems must require negligible human labor.
2. **Safe**: A safe system applies its efforts in the way that its users would want. In particular, if it makes money or acquires other resources or influence, the user retains effective control over most of those resources.

(Note that a system may involve both humans and machines.)

We can crudely quantify efficiency by the ratio between a system's cost and the cost of the most efficient system that can do the same task. We can crudely quantify safety by the expected fraction of the system's outputs which the user captures (with the remainder being applied towards some other ends).

A simple model is that a system's output depends linearly on its efficiency, and that the output is split between the user and whatever other goals the system is implicitly or explicitly pursuing. For example, we could imagine AIs which pursue task-specific goals which are instrumentally valuable to but imperfectly aligned with its users' goals. Then this system may accomplish the particular tasks to which it is applied, but may also divert some resources to pursuing those task-specific goals in ways that the user would not endorse.

Many futurists expect systems to have safety very close to 0 or 1, and that fits into this model just fine (you could still have intermediate degrees of safety, in light of uncertainty).

There may be many unsafe AI systems that don't fit this simple model at all; I'll have this model in mind throughout this post for concreteness, but I don't think that we should too attached to conclusions that are very sensitive to the model (and I'll try to mostly be agnostic).

Note that we can apply a very similar model to the more familiar principal-agent problem, in which a principal tries to induce an agent to act in the principal's interests, and does so imperfectly. The principal can achieve greater efficiency by passing a larger share of their gains on to the agent, or by compromising and adopting values closer to the agents'. This example can help build some intuitions about the case of unsafe AI, but probably shouldn't be taken too seriously.

Social solutions

If we can't always build systems which are efficient and safe, then users must make a tradeoff between safety and efficiency. Users may opt for unsafe systems for a variety of reasons:

- They are maximizing their own returns, which may involve some compromise of safety.
- They share the values pursued by some efficient but unsafe system, or they value the expansion of intelligence or complexity for its own sake.
- They are making a trade with someone or something that shares the values of some efficient but unsafe system, and is willing to subsidize its use.

If the efficiency gains for unsafe systems are large, then resources invested unsafely will compound faster than resources invested safely. The result is that the (proportional) influence of the original human population will be gradually diminished. The quality of our technical solutions determines how large the gap is.

If we coordinate to ensure that no resources are invested unsafely, then we can address the AI control problem regardless of how big this gap is. The difficulty of this coordination problem, the amount of time available to solve it, the strength of the incentives to defect, the necessary level of success, and the consequences of failure, all depend on how good a technical solution to the control problem is available. As the efficiency gap becomes large, all of these parameters become extremely unfavorable.

If the productivity gap is relatively large, say an order of magnitude, then any investment in unsafe systems would very rapidly outpace society at large, and tight controls would be needed to prevent trouble. To put this in quantitative perspective, a 10x productivity boost in the current world would correspond to 30x–1,000x increase in proportional influence per decade (depending on whether the is 10x is compared to average growth or compared to rates of return on other capital investments). So even a very small fraction of unsafe investment could quickly become a major global influence, unless there was a strong social response to thwart the basic economic dynamic.

Examples of social solutions

This is a very broad category. For example, it includes:

- Coordination amongst AI researchers to preferentially develop and distribute safe AI.
- International political coordination to restrict deployment of unsafe AI, or to expropriate resources controlled by unsafe AI and sympathetic humans.
- AI is mostly developed by a single, especially safety-conscious, project. This project maintains a large enough lead over its rivals that it can afford to use inefficient but safe AI, and it manages to preserve this lead either indefinitely or until safe-and-efficient AI can be developed.

Of course there are no clear lines between these categories, or between them and other possible social solutions. I should emphasize

that I am skeptical of most of these approaches, especially the kinds of draconian approaches that could deal with very large efficiency gaps.

I suspect that there will be at least some informal coordination amongst AI researchers, if and when the social benefits of coordination becomes clear. For example, I expect researchers and engineers to generally be happier to work for projects which they see as contributing to human welfare, and this will make life marginally harder for unsafe projects. As a result, the state of the art will be somewhat better for safe systems, it will be somewhat easier for people to apply safe systems to their problems, and so on. I expect this kind of informal coordination to cover small gaps in efficiency; for large efficiency gaps, I could see things going either way.

It would be surprising to me if international political coordination were strong enough to block the adoption of unsafe AI's if they were many times more efficient than their safe counterparts (compare to nuclear disarmament or other international coordination problems we face today). If they were only modestly more efficient, then it would not necessarily require international coordination (just action by the most efficient jurisdictions), and regulation would only need to make very slightly harder for unsafe projects.

I am somewhat skeptical about the feasibility or desirability of world-takeover by an early AI project, though I think that there are exotic situations where it becomes plausible.

Upshot

I suspect that researchers working on the AI control problem should aim high. I think that there is a good chance that society can handle a 10% efficiency gap between safe and unsafe systems, but that it is pretty unlikely that it can handle a 10x gap.

So we capture more value by moving from 10% efficiency to 90% efficiency for safe systems, than by moving from 1% to 10% or from 0.001% to 1%. The relative benefits are mostly determined by the probability that social solutions will be good enough to handle gaps of each size.

On the flip side, I don't think that we need to focus on reaching 100% efficiency either. 90% efficiency seems high enough that there is a good chance that social solutions can handle the gap.

I also think that there are meaningful benefits from progress on both technical and social solutions. I don't think that social solutions are so robust that we don't have to worry about having a very good technical solution (I'm surprised by how often I encounter this view!), nor do I think that technical solutions are so clearly bimodal that we don't care about social solutions since we will be either doomed or have no problem at all.

Postscript: "Foom"

I think that the breakdown above applies whether you expect AI development to proceed briskly or slowly.

If we expect AI development to go really fast (think days or weeks to go from "who cares" to "radically superhuman"), while the rest of the world continues to operate on its comparatively glacial pace, then it becomes more plausible that the first developers of AI will take over the world before anyone else catches up. This provides an easy social solution, and generally lowers the bar for technical solutions: maybe it's OK if their AI is 10x less efficient than it would otherwise be, so that it takes it 10 weeks to take over the world rather than 1 week. This is only a problem if the competition is right on their tails.

I don't think this situation is very likely. But if you do, then you should become more interested in going from 0.001% to 10% efficiency, since that may be all that you really need. You are also probably working on a somewhat different set of safety problems (and I suspect are overall more pessimistic).