Handling destructive technology

Paul Christiano Follow

Nov 14, 2016 · 5 min read

Some technologies are much more helpful for destroying stuff than protecting it. If Alice and Bob both have large nuclear arsenals, and one of them wants the world destroyed, then the world is probably going to get destroyed. If everyone has a large nuclear arsenal and *anyone* wants the world destroyed—well, so it goes.

Preventing nuclear weapons from causing incredible destruction required some global coordination and luck. But things could have been a lot worse; imagine a universe where running twenty amps through a solenoid is enough to destroy the world.

It seems likely that we will eventually encounter technologies that are much more problematic than nuclear weapons: that are much cheaper and more destructive, that don't require hard-to-obtain materials, or that have larger economic importance with fewer close substitutes. Coping with such technologies would require much better global coordination. In the limiting case, the use of sophisticated technologies might need to be centrally coordinated, such that no agent could decide to use them in a destructive way.

In the context of this post, I'll refer to this hypothetical situation as civilization "having our house in order."

AI and delaying the inevitable

If we are not able to solve the control problem, I think that AI will itself be a destructive technology. The incentives to use powerful AI will be significant and the world will move increasingly rapidly and incomprehensibly, but it will be difficult to build competent AI systems that represent human interests. Human values will eventually fade from prominence, just as surely as if humanity had destroyed itself in a nuclear war.

This motivates me to work on AI control, and I think it is amongst the most important technical problems that we can work on today.

But whether or not AI itself is a destructive technology, it seems likely that we will eventually encounter much more destructive physical technologies, and we will need to have our house in order at that point.

So solving the AI control problem merely delays the inevitable. If we get our house in order *prior* to the development of AI then it won’t matter whether we solved the control problem in advance. And if we *never* get our house in order, then it doesn’t matter whether or not we solved the control problem, we are doomed anyway. AI control only matters in the case where we get our house in order between the development of AI and the next time we develop a comparably destructive technology.

I think that solving the control problem is very important anyway, even if it only buys us time:

- Buying time straightforwardly improves our odds of solving the problem by giving us at least one extra chance.
- As technology improves I think we have a better and better shot at getting our house in order. And there are other trends that seem to point in a positive direction as well.
- I think that AI itself may be an important part of getting our house in order, both by expanding the human capability for reason and by opening up new coordination mechanisms.
- There is at least a chance that there won’t be any really destructive technologies after AI, or that such technologies won’t occur again for a very long time. Other changes like independence from earth’s ecology and dispersion to space might be adequate to protect us from upcoming physical technologies (though they clearly wouldn’t be sufficient to protect us from failure to solve the control problem).
- It is possible that the transition to AI will be a natural moment to get our house in order, e.g. because it results in a disruption of the existing world order. I feel hesitant about this possibility, and think that if the development of AI involves an “upset of the existing world order” we are probably in for a bad time. But it’s definitely a possibility.

Overall, it is correct to claim that AI control is less important because we we will need to get our house in order eventually anyway. But I don’t think it changes the cost-benefit analysis very much—I think you’d have a hard time arguing for a factor of 2, much less a factor of 10.

Delaying the inevitable how far?

I expect that sophisticated AI will significantly increase the pace of technological progress, just as the pace of change today is much faster than it was 300 years ago, which was in turn much faster than the pace of change 3000 years ago.

If we imagine a future where major technological change occurs over the scale of years or months rather than decades, it may feel like we aren't going to "delay the inevitable" very far. On this view, if we don't have our house in order by the time we develop sophisticated AI, then we aren't going to have time to get it in order before new destructive technologies are developed.

I encounter this view all of the time, but I think it is probably mistaken. This argument implicitly measures developments by *calendar time*—how many years elapsed between the development of AI and the development of destructive physical technology? If we haven't gotten our house in order by 2045, goes the argument, then what chance do we have of getting our house in order by 2047?

But in the worlds where AI radically increases the pace of technological progress, this is the wrong way to measure. In those worlds science isn't being done by humans, it is being done by a complex ecology of interacting machines moving an order of magnitude faster than modern society. Probably it's not just science: everything is getting done by a complex ecology of interacting machines at unprecedented speed.

If we want to ask about "how much stuff will happen", or "how much change we will see", it is more appropriate to think about *subjective time*: how much thinking and acting actually got done? It doesn't really matter how many times the earth went around the sun.

Now it might be that AI speeds up technological progress without speeding up the process of getting our house in order. My best guess is that if we actually solve the control problem, then it's actually the other way around; but I certainly agree that both are possible.

Even if we assigned only a 50% chance of AI accelerating the process of getting-our-house-in-order as much as it accelerates technological progress, that would cut the value of the control problem by at most a factor of 2. I don't think this is enough to really change the basic calculus for anyone considering whether to invest resources in AI control.

Conclusion

In some sense solving the AI control problem is merely “delaying the inevitable,” pushing back the moment when society needs to have its house in good enough order that it can cope with the discovery of powerful destructive technologies.

I think that this observation can help clarify exactly what we are doing when we work on AI control, and may change how we think about other interventions to safeguard humanity’s future. But I don’t think that it substantially changes the cost-benefit analysis for AI control itself.

I often encounter the argument that AI will facilitate access to powerful destructive technologies, and for that reason requires us to have our house in order whether or not we solve the control problem. I think that this argument is mistaken, most of all by overlooking the ability of AI itself to help us get our house in order.