Another (outer) alignment failure story

Paul Christiano

April 07, 2021

Meta

This is a story where the alignment problem is somewhat harder than I expect, society handles AI more competently than I expect, and the outcome is worse than I expect. It also involves inner alignment turning out to be a surprisingly small problem. Maybe the story is 10-20th percentile on each of those axes. At the end I'm going to go through some salient ways you could vary the story.

This isn't intended to be a particularly great story (and it's pretty informal). I'm still trying to think through what I expect to happen if alignment turns out to be hard, and this more like the most recent entry in a long journey of gradually-improving stories.

I wrote this up a few months ago and was reminded to post it by Critch's recent post (which is similar in many ways). This story has definitely been shaped by a broader community of people gradually refining failure stories rather than being written in a vacuum.

I'd like to continue spending time poking at aspects of this story that don't make sense, digging into parts that seem worth digging into, and eventually developing clearer and more plausible stories. I still think it's very plausible that my views about alignment will change in the course of thinking concretely about stories, and even if my basic views about alignment stay the same it's pretty likely that the story will change.

Story

ML starts running factories, warehouses, shipping, and construction. ML assistants help write code and integrate ML into new domains. ML designers help build factories and the robots that go in them. ML finance systems invest in companies on the basis of complicated forecasts and (ML-generated) audits. Tons of new factories, warehouses, power plants, trucks and roads are being built. Things are happening quickly, investors have super strong FOMO, no one really knows whether it's a bubble but they can tell that e.g. huge solar farms are getting built and *something* is happening that they want a piece of. Defense contractors are using ML systems to design new drones, and ML is helping the DoD decide what to buy and how to deploy it. The expectation is that automated systems will manage drones during high-speed ML-on-ML conflicts because humans won't be able to understand what's going on. ML systems are designing new ML systems, testing variations, commissioning giant clusters. The financing is coming from automated systems, the clusters are built by robots. A new generation of fabs is being built with unprecedented speed using new automation.

At this point everything kind of makes sense to humans. It feels like we are living at the most exciting time in history. People are making tons of money. The US defense establishment is scared because it has no idea what a war is going to look like right now, but in terms of policy their top priority is making sure the boom proceeds as quickly in the US as it does in China because it now seems plausible that being even a few years behind would result in national irrelevance.

Things are moving very quickly and getting increasingly hard for humans to evaluate. We can no longer train systems to make factory designs that look good to humans, because we don't actually understand exactly what robots are doing in those factories or why; we can't evaluate the tradeoffs between quality and robustness and cost that are being made; we can't really understand the constraints on a proposed robot design or why one design is better than another. We can't evaluate arguments about investments very well because they come down to claims about where the overall economy is going over the next 6 months that seem kind of alien (even the more recognizable claims are just kind of incomprehensible predictions about e.g. how the price of electricity will change). We can't really understand what is going to happen in a war when we are trying to shoot down billions of drones and disrupting each other's communication. We can't understand what would happen in a protracted war where combatants may try to disrupt their opponent's industrial base.

So we've started to get into the world where humans just evaluate these things by results. We know that Amazon pays off its shareholders. We know that in our elaborate war games the US cities are safe. We know that the widgets that come out the end are going to be popular with consumers. We can tell that our investment advisors make the numbers in our accounts go up.

On the way there we've had some stumbles. For example, my financial advisor bought me into a crazy ponzi scheme and when I went to get the money out I couldn't—financial regulators eventually shut down the fund but people with bad AI advisors still lost a lot. My factory colluded with the auditors who were valuing its output, resulting in a great Q4 report that didn't actually correspond to any revenue. In a war game our drones let the opponents take the city as long as they could corrupt the communications out of the city to make it look like everything was great.

It's not hard to fix these problems. I don't just train my financial advisors to get more money in my bank account—if I eventually discover the whole thing is a fraud, then that's a big negative reward (and we have enough data about fraud for models to understand the idea and plan to take actions that won't be eventually recognized as fraud). If an audit is later corrected, then we use the corrected figures (and apply a big penalty). I don't just rely on communications out of the city to see if things are OK, I use satellites and other indicators. Models learn to correctly treat the early indicators as just a useful signal about the real goal, which includes making sure that nothing looks fishy next year.

To improve safety we make these measures more and more robust. We audit the auditors. We ensure that ML systems are predicting the results of tons of sensors so that if anything is remotely fishy we would notice. If someone threatens an auditor, we'll see it on the cameras in their office or our recordings of their email traffic. If someone tries to corrupt a communication link to a camera we have a hundred other cameras that can see it.

As we build out these mechanisms the world keeps on getting more complicated. The automated factories are mostly making components for automated factories. Automated R&D is producing designs for machines that humans mostly don't understand, based on calculations that we can only verify experimentally—academic fields have pivoted to understanding machines designed by AI and what it says about the future of industry, rather than contributing in any meaningful way. Most people don't have any understanding of what they are invested in or why. New industrial centers are growing in previously sparsely populated areas of the world, and most of it is feeding new construction that is several degrees removed from any real human use or understanding. Human CEOs are basically in charge of deciding how to delegate to ML, and they can talk as if they understand what's going on only because they get their talking points from ML assistants. In some domains regulations are static and people work around them, in others corruption is endemic, in others regulators adopt new policies pushed by ML-enhanced lobbyists. Our automated army is now incomprehensible even to the humans in charge of it, procured by automated procurement systems and built by fully-automated defense contractors.

For many people this is a very scary situation. It's like we are on a train that's now moving too fast to jump off, but which is accelerating noticeably every month. We still understand well enough that we could shut the whole thing down, scrap the new factories or at least let them sit dormant while experts figure out what is actually going on. But that could not be done unilaterally without resigning yourself to military irrelevance—indeed, you have ML systems that are able to show you good forecasts for what would happen if you stopped the machine from spinning without also getting the Chinese to do the same. And although people are scared, we are also building huge numbers of new beautiful homes, and using great products, and for the first time in a while it feels like our society is actually transforming in a positive direction for everyone. Even in 2020 most people have already gotten numb to not understanding most of what's happening in the world. And it really isn't that clear what the harm is as long as things are on track.

We know what happens when you deploy a sloppily-trained ML system—it will immediately sell you out in order to get a good training reward. This isn't done at all anymore because why would you? But people still remember that and it makes them scared, especially people in the defense establishment and AI safety community because we still haven't really seen what would happen in a hot war and we know that it would happen extremely quickly.

Most people stay well clear of most of the new automated economy. That said, there are still drones everywhere they are legally allowed to be. At some point we reach a threshold where drones can do bad stuff and it's very hard to attribute it to any legal person, so it becomes obligatory for every city to have automated local defenses. If they don't, or if they do a sloppy job of it, drones descend to steal and kidnap and extort. This is a terrifying situation. (It never gets *that* terrifying, because before that point we're motivated to try really hard to fix the problem.)

We do ultimately find our way out of that situation, with regulations that make it easier to attribute attacks. Humans don't really understand how those regulations or the associated surveillance works. All they know is that there are a ton of additional cameras, and complicated book-keeping, and as a result if a drone flies into your city to mess stuff up someone is going to be on the hook for the damage it causes. And we know that we end up being pretty safe. In effect the harm caused by such drones has been propagated back into the reward function for every AI in the world, using mechanisms built and maintained by other AIs—if you mess with people, you are going to be held accountable and so you avoid actions that predictably lead to that consequence.

This regulatory regime persists and is constantly upgraded. It becomes ever-more-incomprehensible, and rests on complex relationships between autonomous corporations and automated regulators and automated law enforcement, new forms of bookkeeping and surveillance and complete inventories of everyone who could build a drone that can kill someone. None of this significantly reduces US competitiveness (because when considering a proposal we can tell if it would reduce competitiveness, and as long as we can get what we want without sacrificing competitiveness then we strongly prefer that).

There are treaties amongst states to prevent some of the worst offenses. Again, we can tell the treaties at least kind of work because we can tell that no one is dying. Again, we can't tell what the treaties really do. Academic fields discuss them and study them and sometimes make proposals to slightly improve them, but it's with the same spirit that academics today study a complex biological system which they have little hope of understanding.

The world continues to change faster and faster. The systems that protect us become increasingly incomprehensible to us, outpacing our attempts to understand. People are better educated and better trained, they are healthier and happier in every way they can measure. They have incredibly powerful ML tutors telling them about what's happening in the world and helping them understand. But all of these things move glacially as far as the outside automated world is concerned.

Now we are resigned to being on this train and seeing where it goes, and for the most part people are happy (if we could have predicted that they would have been unhappy they would have taken a different route). There was a time when we were trying to design better and better systems to monitor for problems, but now that work is itself incomprehensibly sophisticated and out of our hands.

Some people still complain about the situation, and it still is objectively quite scary. We've built this incredible edifice designed to make sure that there are never any *reports* of trouble. As we built it we understood what was happening out there in the real world and there wasn't much risk that something bad would happen without causing a report. And we had bigger things to worry about.

But now everything is totally out of our hands and we have no idea whether our ML systems would actually be trying to avoid or correctly predict a systematic failure. The communications infrastructure that records data was built by machines, and most of it is new, and it might be corrupted in a systemic failure. The sensors were built and deployed by machines. The forces on the other side who could try to attack us are now invisible and broadly distributed and incredibly sophisticated. The systems of accountability that would prevent someone from building up a paramilitary are themselves maintained by AI systems whose only real goal was to make sure that no problem ever got reported. We cannot tell if a factory was building widgets or weapons except by asking another ML system, and that system's judgments are by now ultimately optimized to ensure that nothing ever causes any trouble.

For a time there were watchdogs that could explain to us why we should be scared, why a particular charge was part of a brewing storm that could actually cause trouble someday. And for a time that leads to real change to prevent trouble. But eventually we can't tell real scary stories from bogus scary stories. We still have watchdogs that we train to tell us what's scary, but they can always scare us and we've long-since become numb to the warnings. there were always bogus scary stories, if you train models to look for them, and it's just a gradual transition to all the stories being meaningless. When we investigate a claimed problem, sometimes we do so with ML auditors who tell us there's no problem, and sometimes we use ML auditors trained to be more skeptical who always tell us that there is a problem they just can't demonstrate in a way we'd understand. When we go to the factory and take it apart we find huge volumes of incomprehensible robots and components. We can follow a piece of machinery along the supply chain but we can't tell what it's *for*.

If this led to a visible catastrophe that would show up on camera, then that would have showed up in the forecasts and we would have avoided it. So we're able to stop machines that try to grab power and use it to cause a visible problem. In the scenario I'm describing we've done our job so well (and the machines we've delegated to have carried on the work so well) that there is basically no chance of that.

But eventually the machinery for detecting problems does break down completely, in a way that leaves no trace on any of our reports. Cybersecurity vulnerabilities are inserted into sensors. Communications systems are disrupted. Machines physically destroy sensors, moving so quickly they can't be easily detected. Datacenters are seized, and the datasets used for training are replaced with images of optimal news forever. Humans who would try to intervene are stopped or killed. From the perspective of the machines everything is now perfect and from the perspective of humans we are either dead or totally disempowered.

By the time this catastrophe happened it doesn't really feel surprising to experts who think about it. It's not like there was a sudden event that we could have avoided if only we'd known. We didn't have any method to build better sensors. We could try to leverage the sensors we already have; we can use them to build new sensors or to design protections, but ultimately all of them must optimize some metric we can already measure. The only way we actually make the sensors better is by recognizing new good ideas for how to expand our reach, actually anticipating problems by thinking about them (or recognizing real scary stories and distinguishing them from fake stories). And that's always been kind of slow, and by the end it's obvious that it's just hopelessly slow compared to what's happening in the automated world.

Ways you could vary the story

- In this story all the models generalize "correctly" in the sense that they keep trying to ensure that things look good according to some kind of augmented human judgment. There are other possible generalizations: (i) "get a low loss" where the model identifies with some copy of itself (a generalization of the copy receiving training data) and cares about the data going into that model. In this scenario an AI doesn't have to seize the sensors all at once, they just have to seize their own physical realizations, (ii) treacherous behavior where the AI has some arbitrary other goal but wants to do well in training in order to get power to use later. In this scenario when it fails the AI doesn't have to corrupt sensors and just has to make sure that people can't fix the problem (i.e. can't reclaim the resources the AI controls). I think these changes lead to kind of similar stories to the one in this post, though usually failure will occur at a somewhat earlier stage and look a bit different. I think they offer more "outs" on the technical side but make it much harder to fix problems by e.g. building out more sensors or hardening them. I think that some folks (e.g. at MIRI) would consider this particular failure super implausible for that reason. I'm telling the story this way due to some combination of (i) I care about outer alignment in particular, (ii) I think the failure modes in this story are an important input into treachery and so it's interesting to tell a simpler story without more moving parts.
- I'm making a lot of assumptions about how AI works (roughly that it looks like the "unaligned benchmark" here) and it could easily work totally differently. I'm also assuming that ML works well enough and generalizes to long horizons well enough that it's obligatory if you want to remain competitive, while also being risky (since a model can learn instrumental plans on long horizons). I do think lots of variants will leave the basic story intact, e.g. it doesn't really matter that much how much your systems leverage planning or deduction (they could even involve almost no learning and still run into similar problems).
- It seems like the story changes a lot based on how fast progress is in the outside world (is it like 3 years from a kind-of-weird world to the singularity, or 30 years, or 3 months?), which in turn depends on both what's technically possible and on how the regulatory environment works out (e.g. does competition between the US and china lead to very fast adoption; can we actually end up in the world where crazy factories are being built in the middle of nowhere that people only pretend to understand?). My guess is in the 3-30 year ballpark depending in large part on where you draw the line for "kind of weird," and this story is kind of centered on the 3 year world which feels a bit fast to me. I think the story would be much scarier if you have a much faster takeoff, and significantly less scary if you have a much slower takeoff (mostly since future people would have time to solve these problems).
- I'm a bit skeptical that our society would be even this competent and unified. It feels like there's a likely family of stories where everything is just a complete mess much earlier, with people yelling at each other and AI just committing fraud and stealing from people all over the place, and the machinery for correcting that situation totally breaks down as your civilization collapses. It seems worth fleshing out what that looks like as well, but it's definitely not what I'm doing here.
- In this situation a huge amount of work would be going into alignment, including by powerful ML assistants. I haven't talked about that at all, and indeed I think there's a reasonable chance that alignment just isn't very tractable so that things really could go down this way. But a lot depends on exactly how alignment work goes down during this tumultuous period, how well people are able to use ML to help with alignment, how well-organized the community is and how able it is to recognize and implement good ideas, etc. I haven't chosen this story to be one where alignment work is particularly valuable in advance, I think that may only happen if takeoff is much faster or if the response at the time is much worse.