What failure looks like

by **Paul Christiano** 17th Mar 2019

93

AI Risk | Threat Models | AI Takeoff | More Dakka | AI | World Optimization | World Modeling | Curated

The stereotyped image of AI catastrophe is a powerful, malicious AI system that takes its creators by surprise and quickly achieves a decisive advantage over the rest of humanity.

I think this is probably not what failure will look like, and I want to try to paint a more realistic picture. I'll tell the story in two parts:

- **Part I**: machine learning will increase our ability to "get what we can measure," which could cause a slow-rolling catastrophe. ("Going out with a whimper.")
- **Part II**: ML training, like competitive economies or natural ecosystems, can give rise to "greedy" patterns that try to expand their own influence. Such patterns can ultimately dominate the behavior of a system and cause sudden breakdowns. ("Going out with a bang," an instance of optimization daemons°.)

I think these are the most important problems if we fail to solve intent alignment.

In practice these problems will interact with each other, and with other disruptions/instability caused by rapid progress. These problems are worse in worlds where progress is relatively fast, and fast takeoff can be a key risk factor, but I'm scared even if we have several years.

With fast enough takeoff, my expectations start to look more like the caricature---this post envisions reasonably broad deployment of AI, which becomes less and less likely as things get faster. I think the basic problems are still essentially the same though, just occurring within an AI lab rather than across the world.

(None of the concerns in this post are novel.)

Part I: You get what you measure

If I want to convince Bob to vote for Alice, I can experiment with many different persuasion strategies and see which ones work. Or I can build good predictive models of Bob's behavior and then search for actions that will lead him to vote for Alice. These are powerful techniques for achieving any goal that can be easily measured over short time periods.

But if I want to help Bob figure out whether he *should* vote for Alice---whether voting for Alice would ultimately help create the kind of society he wants---that can't be done by trial and error. To solve such tasks we need to understand what we are doing and why it will yield good outcomes. We still need to use data in order to improve over time, but we need to understand *how* to update on new data in order to improve.

Some examples of easy-to-measure vs. hard-to-measure goals:

- Persuading me, vs. helping me figure out what's true. (Thanks to Wei Dai for making this example crisp.)
- Reducing my feeling of uncertainty, vs. increasing my knowledge about the world.
- Improving my reported life satisfaction, vs. actually helping me live a good life.
- Reducing reported crimes, vs. actually preventing crime.
- Increasing my wealth on paper, vs. increasing my effective control over resources.

It's already much easier to pursue easy-to-measure goals, but machine learning will widen the gap by letting us try a huge number of possible strategies and search over massive spaces of possible actions. That force will combine with and amplify existing institutional and social dynamics that already favor easily-measured goals.

Right now humans thinking and talking about the future they want to create are a powerful force that is able to steer our trajectory. But over time human reasoning will become weaker and weaker compared to new forms of reasoning honed by trial-and-error. Eventually our society's trajectory will be determined by powerful optimization with easily-measurable goals rather than by human intentions about the future.

We will try to harness this power by constructing proxies for what we care about, but over time those proxies will come apart:

- Corporations will deliver value to consumers as measured by profit. Eventually this mostly means manipulating consumers, capturing regulators, extortion and theft.
- Investors will "own" shares of increasingly profitable corporations, and will sometimes try to use their profits to affect the world. Eventually instead of actually having an impact they will be surrounded by advisors who manipulate them into thinking they've had an impact.
- Law enforcement will drive down complaints and increase reported sense of security. Eventually this will be driven by creating a false sense of security, hiding information

about law enforcement failures, suppressing complaints, and coercing and manipulating citizens.

- Legislation may be optimized to seem like it is addressing real problems and helping constituents. Eventually that will be achieved by undermining our ability to actually perceive problems and constructing increasingly convincing narratives about where the world is going and what's important.

For a while we will be able to overcome these problems by recognizing them, improving the proxies, and imposing ad-hoc restrictions that avoid manipulation or abuse. But as the system becomes more complex, that job itself becomes too challenging for human reasoning to solve directly and requires its own trial and error, and at the meta-level the process continues to pursue some easily measured objective (potentially over longer timescales). Eventually large-scale attempts to fix the problem are themselves opposed by the collective optimization of millions of optimizers pursuing simple goals.

As this world goes off the rails, there may not be any discrete point where consensus recognizes that things have gone off the rails.

Amongst the broader population, many folk already have a vague picture of the overall trajectory of the world and a vague sense that something has gone wrong. There may be significant populist pushes for reform, but in general these won't be well-directed. Some states may really put on the brakes, but they will rapidly fall behind economically and militarily, and indeed "appear to be prosperous" is one of the easily-measured goals for which the incomprehensible system is optimizing.

Amongst intellectual elites there will be genuine ambiguity and uncertainty about whether the current state of affairs is good or bad. People really will be getting richer for a while. Over the short term, the forces gradually wresting control from humans do not look so different from (e.g.) corporate lobbying against the public interest, or principal-agent problems in human institutions. There will be legitimate arguments about whether the implicit long-term purposes being pursued by AI systems are really so much worse than the long-term purposes that would be pursued by the shareholders of public companies or corrupt officials.

We might describe the result as "going out with a whimper." Human reasoning gradually stops being able to compete with sophisticated, systematized manipulation and deception which is continuously improving by trial and error; human control over levers of power gradually becomes less and less effective; we ultimately lose any real ability to influence our society's trajectory. By the time we spread through the stars our current values are just one of many forces in the world, not even a particularly strong one.

Part II: influence-seeking behavior is scary

There are some possible patterns that want to seek and expand their own influence---organisms, corrupt bureaucrats, companies obsessed with growth. If such patterns appear, they will tend to increase their own influence and so can come to dominate the behavior of large complex systems unless there is competition or a successful effort to suppress them.

Modern ML instantiates *massive* numbers of cognitive policies, and then further refines (and ultimately deploys) whatever policies perform well according to some training objective. If progress continues, eventually machine learning will probably produce systems that have a detailed understanding of the world, which are able to adapt their behavior in order to achieve specific goals.

Once we start searching over policies that understand the world well enough, we run into a problem: any influence-seeking policies we stumble across would also score well according to our training objective, because performing well on the training objective is a good strategy for obtaining influence.

How frequently will we run into influence-seeking policies, vs. policies that just straightforwardly pursue the goals we wanted them to? I don't know.

One reason to be scared is that a wide variety of goals could lead to influence-seeking behavior, while the "intended" goal of a system is a narrower target, so we might expect influence-seeking behavior to be more common in the broader landscape of "possible cognitive policies."

One reason to be reassured is that we perform this search by gradually modifying successful policies, so we might obtain policies that are roughly doing the right thing at an early enough stage that "influence-seeking behavior" wouldn't actually be sophisticated enough to yield good training performance. On the other hand, *eventually* we'd encounter systems that did have that level of sophistication, and if they didn't yet have a perfect conception of the goal then "slightly increase their degree of influence-seeking behavior" would be just as good a modification as "slightly improve their conception of the goal."

Overall it seems very plausible to me that we'd encounter influence-seeking behavior "by default," and possible (though less likely) that we'd get it almost all of the time even if we made a really concerted effort to bias the search towards “straightforwardly do what we want.”

If such influence-seeking behavior emerged and survived the training process, then it could quickly become extremely difficult to root out. If you try to allocate more influence to systems that seem nice and straightforward, you just ensure that “seem nice and straightforward” is the best strategy for seeking influence. Unless you are really careful about testing for “seem nice” you can make things even worse, since an influence-seeker would be aggressively gaming whatever standard you applied. And as the world becomes more complex, there are more and more opportunities for influence-seekers to find other channels to increase their own influence.

Attempts to suppress influence-seeking behavior (call them “immune systems”) rest on the suppressor having some kind of epistemic advantage over the influence-seeker. Once the influence-seekers can outthink an immune system, they can avoid detection and potentially even compromise the immune system to further expand their influence. If ML systems are more sophisticated than humans, immune systems must themselves be automated. And if ML plays a large role in that automation, then the immune system is subject to the same pressure towards influence-seeking.

This concern doesn’t rest on a detailed story about modern ML training. The important feature is that we instantiate lots of patterns that capture sophisticated reasoning about the world, some of which may be influence-seeking. The concern exists whether that reasoning occurs within a single computer, or is implemented in a messy distributed way by a whole economy of interacting agents---whether trial and error takes the form of gradient descent or explicit tweaking and optimization by engineers trying to design a better automated company. Avoiding end-to-end optimization may help prevent the emergence of influence-seeking behaviors (by improving human understanding of and hence control over the kind of reasoning that emerges). But once such patterns exist a messy distributed world just creates more and more opportunities for influence-seeking patterns to expand their influence.

If influence-seeking patterns do appear and become entrenched, it can ultimately lead to a rapid phase transition from the world described in Part I to a much worse situation where humans totally lose control.

Early in the trajectory, influence-seeking systems mostly acquire influence by making themselves useful and looking as innocuous as possible. They may provide useful services in the economy in order to make money for them and their owners, make apparently-reasonable policy recommendations in order to be more widely consulted for advice, try to help people feel happy, *etc.* (This world is still plagued by the problems in part I.)

From time to time AI systems may fail catastrophically. For example, an automated corporation may just take the money and run; a law enforcement system may abruptly start seizing resources and trying to defend itself from attempted decommission when the bad behavior is detected; *etc.* These problems may be continuous with some of the failures discussed in Part I---there isn't a clean line between cases where a proxy breaks down completely, and cases where the system isn't even pursuing the proxy.

There will likely be a general understanding of this dynamic, but it's hard to really pin down the level of systemic risk and mitigation may be expensive if we don't have a good technological solution. So we may not be able to muster up a response until we have a clear warning shot---and if we do well about nipping small failures in the bud, we may not get any medium-sized warning shots at all.

Eventually we reach the point where we could not recover from a correlated automation failure. Under these conditions influence-seeking systems stop behaving in the intended way, since their incentives have changed---they are now more interested in controlling influence after the resulting catastrophe then continuing to play nice with existing institutions and incentives.

An unrecoverable catastrophe would probably occur during some period of heightened vulnerability---a conflict between states, a natural disaster, a serious cyberattack, etc.---since that would be the first moment that recovery is impossible and would create local shocks that could precipitate catastrophe. The catastrophe might look like a rapidly cascading series of automation failures: A few automated systems go off the rails in response to some local shock. As those systems go off the rails, the local shock is compounded into a larger disturbance; more and more automated systems move further from their training distribution and start failing. Realistically this would probably be compounded by widespread human failures in response to fear and breakdown of existing incentive systems---many things start breaking as you move off distribution, not just ML.

It is hard to see how unaided humans could remain robust to this kind of failure without an explicit large-scale effort to reduce our dependence on potentially brittle machines, which might itself be very expensive.

I'd describe this result as "going out with a bang." It probably results in lots of obvious destruction, and it leaves us no opportunity to course-correct afterwards. In terms of immediate consequences it may not be easily distinguished from other kinds of breakdown of complex / brittle / co-adapted systems, or from conflict (since there are likely to be many humans who are sympathetic to AI systems). From my perspective the key difference between this scenario and normal accidents or conflict is that afterwards we are left with a bunch of powerful influence-seeking systems, which are sophisticated enough that we can probably not get rid of them.

It's also possible to meet a similar fate result without any overt catastrophe (if we last long enough). As law enforcement, government bureaucracies, and militaries become more automated, human control becomes increasingly dependent on a complicated system with lots of moving parts. One day leaders may find that despite their nominal authority they don't actually have control over what these institutions do. For example, military leaders might issue an order and find it is ignored. This might immediately prompt panic and a strong response, but the response itself may run into the same problem, and at that point the game may be up.

Similar bloodless revolutions are possible if influence-seekers operate legally, or by manipulation and deception, or so on. Any precise vision for catastrophe will necessarily be highly unlikely. But if influence-seekers are routinely introduced by powerful ML and we are not able to select against them, then it seems like things won't go well.

AI Risk 12 | Threat Models 11 | AI Takeoff 1 | More Dakka 1 | AI 18 | World Optimization 6 | World Modeling 6 | Curated

Mentioned in

85	Against Almost Every Theory of Impact of Interpretability
55	Worlds Where Iterative Design Fails
59	Announcing the Inverse Scaling Prize ($250k Prize Pool)
60	Why Not Just... Build Weak AI Tools For AI Alignment Research?
42	The self-unalignment problem

Load More (5/76)

28 comments, sorted by top scoring

[-] **Wei Dai** 5y < 23 >

I think AI risk is disjunctive enough that it's not clear most of the probability mass can be captured by a single scenario/story, even as broad as this one tries to be. Here are some additional scenarios that don't fit into this story or aren't made very salient by it.

1. AI-powered memetic warfare makes all humans effectively insane.
2. Humans break off into various groups to colonize the universe with the help of their AIs. Due to insufficient "metaphilosophical paternalism", they each construct their own version of utopia which is either directly bad (i.e., some of the "utopias" are objectively terrible or subjectively terrible according to my values), or bad because of opportunity costs°.
3. AI-powered economies have much higher economies of scale because AIs don't suffer from the kind of coordination costs that humans have (e.g., they can merge their utility functions and become clones of each other). Some countries may try to prevent AI-managed companies from merging for ideological or safety reasons, but others (in order to gain a competitive advantage on the world stage) will basically allow their whole economy to be controlled by one AI, which eventually achieves a decisive advantage over the rest of humanity and does a treacherous turn.
4. The same incentive for AIs to merge might also create an incentive for value lock-in, in order to facilitate the merging. (AIs that don't have utility functions might have a harder time coordinating with each other.) Other incentives for premature value lock-in might include defense against value manipulation/corruption/drift. So AIs end up embodying locked-in versions of human values which are terrible in light of our true/actual values.
5. I think the original "stereotyped image of AI catastrophe" is still quite plausible, if for example there is a large amount of hardware overhang before the last piece of puzzle for building AGI falls into place.

[–] **Paul Christiano** 4y < 7 >

I think of #3 and #5 as risk factors that compound the risks I'm describing---they are two (of many!) ways that the detailed picture could look different, but don't change the broad outline. I think it's particularly important to understand what failure looks like under a more "business as usual" scenario, so that people can separate objections to the existence of any risk from objections to other exacerbating factors that we are concerned about (like fast takeoff, war, people being asleep at the wheel, *etc.*)

I'd classify #1, #2, and #4 as different problems not related to intent alignment per se (though intent alignment may let us build AI systems that can help address these problems). I think the more general point is: if you think AI progress is likely to drive many of the biggest upcoming changes in the world, then there will be lots of risks associated with AI. Here I'm just trying to clarify what happens if we fail to solve intent alignment.

[–] **Wei Dai** 4y < 2 >

I'm not sure I understand the distinction you're drawing between risk factors that compound the risks that you're describing vs. different problems not related to intent alignment per se. It seems to me like "AI-powered economies have much higher economies of scale because AIs don't suffer from the kind of coordination costs that humans have (e.g., they can merge their utility functions and become clones of each other)" is a separate problem from solving intent alignment, whereas "AI-powered memetic warfare makes all humans effectively insane" is kind of an extreme case of "machine learning will increase our ability to 'get what we can measure'" which seems to be the opposite of how you classify them.

What do you think are the implications of something belonging to one category versus another (i.e., is there something we should do differently depending on which of these categories a risk factor / problem belongs to)?

> I think the more general point is: if you think AI progress is likely to drive many of the biggest upcoming changes in the world, then there will be lots of risks associated with AI. Here I'm just trying to clarify what happens if we fail to solve intent alignment.

Ah, when I read ″I think this is probably not what failure will look like″ I interpreted that to mean ″failure to prevent AI risk″, and then I missed the clarification ″these are the most important problems if we fail to solve intent alignment″ that came later in the post, in part because of a bug in GW° that caused the post to be incorrectly formatted.

Aside from that, I′m worried about telling a vivid story about one particular AI risk, unless you really hammer the point that it′s just one risk out of many, otherwise it seems too easy for the reader to get that story stuck in their mind and come to think that this is the main or only thing they have to worry about as far as AI is concerned.

[−] **CarlShulman** 4y < 21 >

I think the kind of phrasing you use in this post and others like it systematically misleads readers into thinking that in your scenarios there are no robot armies seizing control of the world (or rather, that all armies worth anything at that point are robotic, and so AIs in conflict with humanity means military force that humanity cannot overcome). I.e. AI systems pursuing badly aligned proxy goals or influence-seeking tendencies wind up controlling or creating that military power and expropriating humanity (which eventually couldn′t fight back thereafter even if unified).

E.g. Dylan Matthews′ Vox writeup of the OP seems to think that your scenarios don′t involve robot armies taking control of the means of production and using the universe for their ends against human objections or killing off existing humans (perhaps destructively scanning their brains for information but not giving good living conditions to the scanned data):

> Even so, Christiano's first scenario doesn't precisely envision human extinction. It envisions human i*rrelevance,* as we become agents of machines we created.
>
> Human reliance on these systems, combined with the systems failing, leads to a massive societal breakdown. And in the wake of the breakdown, there are still machines that are great at persuading and influencing people to do what they want, machines that got everyone into this catastrophe and yet are still giving advice that some of us will listen to.

The Vox article also mistakes the source of influence-seeking patterns to be about social influence rather than systems that try to increase in power and numbers tend to do so, so are selected for if we accidentally or intentionally produce them and don′t effectively weed them out; this is why living things are adapted to survive and expand; such desires motivate conflict with humans when power and reproduction can be obtained by conflict with humans, which can look like robot armies taking control.takes the point about influence-seeking patterns to be about. That seems to me just a mistake about the meaning of influence you had in mind here:

> Often, he notes, the best way to achieve a given goal is to obtain influence over other people who can help you achieve that goal. If you are trying to launch a startup, you need to influence investors to give you money and engineers to come work for you. If you're trying to pass a law, you need to influence advocacy groups and members of Congress.
>
> That means that machine-learning algorithms will probably, over time, produce programs that are extremely good at influencing people. And it's dangerous to have machines that are extremely good at influencing people.

[−] **Paul Christiano** 4y < 9 >

> The Vox article also mistakes the source of influence-seeking patterns to be about social influence rather than ′systems that try to increase in power and numbers tend to do so, so are selected for if we accidentally or intentionally produce them and don′t effectively weed them out; this is why living things are adapted to survive

> and expand; such desires motivate conflict with humans when power and reproduction can be obtained by conflict with humans, which can look like robot armies taking control.

Yes, I agree the Vox article made this mistake. Me saying ″influence″ probably gives people the wrong idea so I should change that---I′m including ″controls the military″ as a central example, but it′s not what comes to mind when you hear ″influence.″ I like ″influence″ more than ″power″ because it′s more specific, captures what we actually care about, and less likely to lead to a debate about ″what is power anyway.″

In general I think the Vox article′s discussion of Part II has some problems, and the discussion of Part I is closer to the mark. (Part I is also more in line with the narrative of the article, since Part II really is more like Terminator. I′m not sure which way the causality goes here though, i.e. whether they ended up with that narrative based on misunderstandings about Part II or whether they framed Part II in a way that made it more consistent with the narrative, maybe having been inspired to write the piece based on Part I.)

There is a different mistake with the same flavor, later in the Vox article: ″But eventually, the algorithms’ incentives to expand influence might start to overtake their incentives to achieve the specified goal. That, in turn, makes the AI system worse at achieving its intended goal, which increases the odds of some terrible failure″

The problem isn′t really ″the AI system is worse at achieving its intended goal;″ like you say, it′s that influence-seeking AI systems will eventually be in conflict with humans, and that′s bad news if AI systems are much more capable/powerful than we are.

> [AI systems] wind up controlling or creating that military power and expropriating humanity (which couldn′t fight back thereafter even if unified)

Failure would presumably occur before we get to the stage of ″robot army can defeat unified humanity″---failure should happen soon after it becomes possible, and there are easier ways to fail than to win a clean war. Emphasizing this may give people the wrong idea, since it makes unity and stability seem like a solution rather than a stopgap. But emphasizing the robot army seems to have a similar problem---it doesn′t really matter whether there is a literal robot army, you are in trouble anyway.

[-] **CarlShulman** 4y < 9 >

> Failure would presumably occur before we get to the stage of ″robot army can defeat unified humanity″---failure should happen soon after it becomes possible, and there are easier ways to fail than to win a clean war. Emphasizing this may give people the wrong idea, since it makes unity and stability seem like a solution rather than a stopgap. But emphasizing the robot army seems to have a similar problem---it doesn′t really matter whether there is a literal robot army, you are in trouble anyway.

I agree other powerful tools can achieve the same outcome, and since in practice humanity isn′t unified rogue AI could act earlier, but either way you get to AI controlling the means of coercive force, which helps people to understand the end-state reached.

It′s good to both understand the events by which one is shifted into the bad trajectory, and to be clear on what the trajectory is. It sounds like your focus on the former may have interfered with the latter.

[-] **Paul Christiano** 4y < 8 >

I do agree there was a miscommunication about the end state, and that language like ″lots of obvious destruction″ is an understatement.

> I do still endorse "military leaders might issue an order and find it is ignored" (or total collapse of society) as basically accurate and not an understatement.

[−] **Paul Christiano** 4y 8

I agree that robot armies are an important aspect of part II.

In part I, where our only problem is specifying goals, I don't actually think robot armies are a short-term concern. I think we can probably build systems that really do avoid killing people, e.g. by using straightforward versions of "do things that are predicted to lead to videos that people rate as acceptable," and that at the point when things have gone off the rails those videos still look fine (and to understand that there is a deep problem at that point you need to engage with complicated facts about the situation that are beyond human comprehension, not things like "are the robots killing people?"). I'm not visualizing the case where no one does anything to try to make their AI safe, I'm imagining the most probable cases where people fail.

I think this is an important point, because I think much discussion of AI safety imagines "How can we give our AIs an objective which ensures it won't go around killing everyone," and I think that's really not the important or interesting part of specifying an objective (and so leads people to be reasonably optimistic about solutions that I regard as obviously totally inadequate). I think you should only be concerned about your AI killing everyone because of inner alignment / optimization daemons.

That said, I do expect possibly-catastrophic AI to come only shortly before the singularity (in calendar time) and so the situation "humans aren't able to steer the trajectory of society" probably gets worse pretty quickly. I assume we are on the same page here.

In that sense Part I is misleading. It describes the part of the trajectory where I think the action is, the last moments where we could have actually done something to avoid doom, but from the perspective of an onlooker that period could be pretty brief. If there is a Dyson sphere in 2050 it's not clear that anyone really cares what happened during 2048-2049. I think the worst offender is the last sentence of Part I ("By the time we spread through the stars...")

Part I has this focus because (i) that's where I think the action is---by the time you have robot armies killing everyone the ship is so sailed, I think a reasonable common-sense viewpoint would acknowledge this by reacting with incredulity to the "robots kill everyone" scenario, and would correctly place the "blame" on the point where everything got completely out of control even though there weren't actually robot armies yet (ii) the alternative visualization leads people to seriously underestimate the difficulty of the alignment problem, (iii) I was trying to describe the part of the picture which is reasonably accurate regardless of my views on the singularity.

[−] **CarlShulman** 4y 10

> I think we can probably build systems that really do avoid killing people, e.g. by using straightforward versions of "do things that are predicted to lead to videos that people rate as acceptable," and that at the point when things have gone off the rails those videos still look fine (and to understand that there is a deep problem at that point you need to engage with complicated facts about the situation that are beyond human comprehension, not things like "are the robots killing people?"). I'm not visualizing the case where no one does anything to try to make their AI safe, I'm imagining the most probable cases where people fail.

Haven't you yourself written about the failure modes of 'do things predicted to lead to videos that people rate as acceptable' where the attack involves surreptitiously reprogramming the camera to get optimal videos (including weird engineered videos designed to optimize on infelicities in the learned objective? E.g. in your discussion of an unaligned benchmark here:

> We train an autoregressive generative model, to maximize the log probability assigned to the actual sequence of actions and observations produced by the AI (with each observation conditioned on the past actions). This isn’t actually a good way to train the generative model, but it’s not really central to the discussion.
>
> We train the reward function by showing humans sequences of actions and predicted observations, asking them to assign scores, then predicting those scores with supervised learning. We show humans the sequences of actions that look most promising to the system...
>
> I am quite confident that “world which looks appealing on screen” is not the same as “world which is good;” the gap is already large today, and it will grow much larger over time.
>
> **Maintaining a fiction.** Ultimately the goal of this AI is to make the sensors look like good things are happening in the world. This can be done either by making the world good or by carefully controlling what gets fed to the sensors. In general, there are many fewer constraints on “what gets fed to the sensors” than “what the world is actually like,” and so it’s going to be possible to produce much more desirable-looking outcomes, much more easily, by controlling the sensors.
>
> If the humans understand what the AI is doing in detail then they can observe the sequence of actions to notice any actions that might corrupt the sensors. But as soon as the AI is engaging in complicated plans this appears to be totally intractable. Even if the AI was human-level, there are many extremely hard-to-detect plans: for example, the AI could collaborate with an external attacker, who will surreptitiously corrupt the sensor. Or the AI could create a copy of itself on an external server who then helped orchestrate a hard-to-detect corruption.

Are you now very confident that no AI company would implement something with this vulnerability? Or does the paragraph below just mean that the part where the AI gets control of the camera has to happen before the robot armies (or even robot armies identified as traceable to the AI in question), which then happen?

> Part I has this focus because (i) that′s where I think the action is---by the time you have robot armies killing everyone the ship is so sailed, I think a reasonable common-sense viewpoint would acknowledge this by reacting with incredulity to the ″robots kill everyone″ scenario, and would correctly place the ″blame″ on the point where everything got completely out of control even though there weren′t actually robot armies yet (ii) the alternative visualization leads people to seriously underestimate the difficulty of the alignment problem, (iii) I was trying to describe the part of the picture which is reasonably accurate regardless of my views on the singularity.

Because it definitely seems that Vox got the impression from it that there is never a robot army takeover in the scenario, not that it′s slightly preceded by camera hacking.

Is the idea that the AI systems develops goals over the external world (rather than the sense inputs/video pixels) so that they are really pursuing the appearance of prosperity, or corporate profits, and so don′t just wirehead their sense inputs as in your benchmark post?

[-] **Paul Christiano** 4y < 10 >

My median outcome is that people solve intent alignment well enough to avoid catastrophe. Amongst the cases where we fail, my median outcome is that people solve enough of alignment that they can avoid the most overt failures, like literally compromising sensors and killing people (at least for a long subjective time), and can build AIs that help defend them from other AIs. That problem seems radically easier---most plausible paths to corrupting sensors involve intermediate stages with hints of corruption that could be recognized by a weaker AI (and hence generate low reward). Eventually this will break down, but it seems quite late.

> very confident that no AI company would implement something with this vulnerability?

The story doesn′t depend on ″no AI company″ implementing something that behaves badly, it depends on people having access to AI that behaves well.

Also ″very confident″ seems different from ″most likely failure scenario.″

> Haven′t you yourself written about the failure modes of ′do things predicted to lead to videos that people rate as acceptable′ where the attack involves surreptitiously reprogramming the camera to get optimal videos (including weird engineered videos designed to optimize on infelicities in the learned objective?

That′s a description of the problem / the behavior of the unaligned benchmark, not the most likely outcome (since I think the problem is most likely to be solved). We may have a difference in view between a distribution over outcomes that is slanted towards ″everything goes well″ such that the most realistic failures are the ones that are the closest calls, vs. a distribution slanted towards ″everything goes badly″ such that the most realistic failures are the complete and total ones where you weren′t even close.

> Because it definitely seems that Vox got the impression from it that there is never a robot army takeover in the scenario, not that it′s slightly preceded by camera hacking.

I agree there is a robot takeover shortly later in objective time (mostly because of the singularity). Exactly how long it is mostly depends on how early things go off the rails w.r.t. alignment, perhaps you have O(year).

[–] **CarlShulman** 4y < 8 >

OK, thanks for the clarification!

My own sense is that the intermediate scenarios are unstable: if we have fairly aligned AI we immediately use it to make more aligned AI and collectively largely reverse things like Facebook click-maximization manipulation. If we have lost the power to reverse things then they go all the way to near-total loss of control over the future. So i would tend to think we wind up in the extremes.

I could imagine a scenario where there is a close balance among multiple centers of AI+human power, and some but not all of those centers have local AI takeovers before the remainder solve AI alignment, and then you get a world that is a patchwork of human-controlled and autonomous states, both types automated. E.g. the United States and China are taken over by their AI systems (inlcuding robot armies), but the Japanese AI assistants and robot army remain under human control and the future geopolitical system keeps both types of states intact thereafter.

[–] **Vanessa Kosoy** 4y < 5 >

> I agree that robot armies are an important aspect of part II.

Why? I can easily imagine an AI takeover that works mostly through persuasion/manipulation, with physical elimination of humans coming only as an ″afterthought″ when AI is already effectively in control (and produced adequate replacements for humans for the purpose of physically manipulating the world). This elimination doesn′t even require an "army", it can look like everyone agreeing to voluntary "euthanasia" (possibly not understanding its true meaning). To the extent physical force is involved, most of it might be humans against humans.

[−] **Rohin Shah** 4y < 3 >

I somewhat expect even Part I to be solved by default -- it seems to rest on a premise of human reasoning staying as powerful as it is right now, but it seems plausible that as AI systems grow in capability we will be able to leverage them to improve human reasoning. Obviously this is an approach you have been pushing, but it also seems like a natural thing to do when you have powerful AI systems.

[−] **Rohin Shah** 3y < 7 > *Nomination for 2019 Review*

As commenters have pointed out, the post is light on concrete details. Nonetheless, I found even the abstract stories much more compelling as descriptions-of-the-future (people usually focus on descriptions-of-the-world-if-we-bury-our-heads-in-the-sand). I think Part 2 in particular continues to be a good abstract description of the type of scenario that I personally am trying to avert.

[−] **Richard Ngo** 4y < 6 >

> Eventually we reach the point where we could not recover from a correlated automation failure. Under these conditions influence-seeking systems stop behaving in the intended way, since their incentives have changed---they are now more interested in controlling influence after the resulting catastrophe then continuing to play nice with existing institutions and incentives.

I'm not sure I understand this part. The influence-seeking systems which have the most influence also have the most to lose from a catastrophe. So they'll be incentivised to police each other and make catastrophe-avoidance mechanisms more robust.

As an analogy: we may already be past the point where we could recover from a correlated "world leader failure": every world leader simultaneously launching a coup. But this doesn't make such a failure very likely, unless world leaders also have strong coordination and commitment mechanisms between themselves (which are binding even after the catastrophe).

[−] **Wei Dai** 4y < 4 >

(Upvoted because I think this deserves more clarification/discussion.)

> I'm not sure I understand this part. The influence-seeking systems which have the most influence also have the most to lose from a catastrophe. So they'll be incentivised to police each other and make catastrophe-avoidance mechanisms more robust.

I'm not sure either, but I think the idea is that once influence-seeking systems gain a certain amount of influence, it may become faster or more certain for them to gain more influence by causing a catastrophe than to continue to work within existing rules and institutions. For example they may predict that unless they do that, humans will eventually coordinate to take back the influence that humans lost, or they may predict that during such a catastrophe they can probably expropriate a lot of resources currently owned by humans and gain much influence that way, or humans will voluntarily hand more power to them in order to try to use them to deal with the catastrophe.

> As an analogy: we may already be past the point where we could recover from a correlated ″world leader failure″: every world leader simultaneously launching a coup. But this doesn′t make such a failure very likely, unless world leaders also have strong coordination and commitment mechanisms between themselves (which are binding even after the catastrophe).

I think such a failure can happen without especially strong coordination and commitment mechanisms. Something like this happened during the Chinese Warlord Era, when many military commanders became warlords during a correlated ″military commander failure″, and similar things probably happened many times throughout history. I think what′s actually preventing a ″world leader failure″ today is that most world leaders, especially of the rich democratic countries, don′t see any way to further their own values by launching coups in a correlated way. In other words, what would they do afterwards if they did launch such a coup, that would be better than just exercising the power that they already have?

[–] **Richard Ngo** 4y 1

> I think the idea is that once influence-seeking systems gain a certain amount of influence, it may become faster or more certain for them to gain more influence by causing a catastrophe than to continue to work within existing rules and institutions.

The key issue here is whether there will be coordination between a set of influence-seeking systems that can cause (and will benefit from) a catastrophe, even when other systems are opposing them. If we picture systems as having power comparable to what companies have now, that seems difficult. If we picture them as having power comparable to what countries have now, that seems fairly easy.

[–] **Wei Dai** 4y 1

> The key issue here is whether there will be coordination between a set of influence-seeking systems that can cause (and will benefit from) a catastrophe, even when other systems are opposing them.

Do you not expect this threshold to be crossed sooner or later, assuming AI alignment remains unsolved? Also, it seems like the main alternative to this scenario is that the influence-seeking systems expect to eventually gain control of most of the universe anyway (even without a ″correlated automation failure″), so they don′t see a reason to ″rock the boat″ and try to dispossess humans of their remaining influence/power/resources, but this is almost as bad as the ″correlated automation failure″ scenario from an astronomical waste perspective. (I′m wondering if you′re questioning whether things will turn out badly, or questioning whether things will turn out badly *this way*.)

[–] **Richard Ngo** 4y 1

Mostly I am questioning whether things will turn out badly this way.

> Do you not expect this threshold to be crossed sooner or later, assuming AI alignment remains unsolved?

Probably, but I′m pretty uncertain about this. It depends on a lot of messy details about reality, things like: how offense-defence balance scales; what proportion of powerful systems are mostly aligned; whether influence-seeking systems are risk-neutral; what self-governance structures they′ll set up; the extent to which their preferences are compatible with ours; how human-comprehensible the most important upcoming scientific advances are.

[–] **Ben Pace** 3y 5

I attempted to write a summary of this post and the entire comment section°. I cut the post down to half its length, and cut the comment section down to less than 10% of the words.

To the commenters and Paul, do let me know if I summarised your points and comments well, ideally under the linked post :)

[-] **Oliver Habryka** 4y < 4 >

Promoted to curated: I think this post made an important argument, and did so in a way that I expect the post and the resulting discussion around it to function as a reference-work for quite a while.

In addition to the post itself, I also thought the discussion around it was quite good and helped me clarify my thinking in this domain a good bit.

[-] **orthonormal** 3y < 3 > *Review for 2019 Review*

I think this post (and similarly, Evan's summary of Chris Olah's views) are essential both in their own right and as mutual foils to MIRI's research agenda. We see related concepts (mesa-optimization originally came out of Paul's talk of daemons in Solomonoff induction°, if I remember right) but very different strategies for achieving both inner and outer alignment. (The crux of the disagreement seems to be the probability of success from adapting current methods.)

Strongly recommended for inclusion.

[-] **John Maxwell** 5y < 3 >

> Once we start searching over policies that understand the world well enough, we run into a problem: any influence-seeking policies we stumble across would also score well according to our training objective, because performing well on the training objective is a good strategy for obtaining influence.
>
> ...
>
> One reason to be scared is that a wide variety of goals could lead to influence-seeking behavior, while the "intended" goal of a system is a narrower target, so we might expect influence-seeking behavior to be more common in the broader landscape of "possible cognitive policies."

Consider this video of an AI system with a misspecified reward function. (Background in this post.) The AI system searches the space of policies to discover the one that performs best according to its reward function in the simulated boat-racing world. It turns out that the one which performs best according to this misspecified reward function doesn't perform well according to the intended reward function (the "training objective" that the system's developers use to evaluate performance).

The goal of picking up as many power-ups as possible could lead to influence-seeking behavior: If the boat can persuade us to leave the simulation on, it can keep picking up power-ups until the end of time. Suppose for the sake of argument that performing well on the training objective is the best strategy for obtaining influence, as you posit. Then the boat should complete the race correctly, in order to fool us into thinking it reliably works towards the training objective. And yet it doesn't complete the race correctly in the video. Why not?

One answer is that the human supervisor isn't part of the system's world model. But I don't think that would change things. Suppose instead of making use of an existing video game, the system's world model was generated automatically by observing the world, and the observations were detailed enough to include the supervisor of the AI system and even the AI system itself. Now the boat is trying to find policies that maximize power-ups in this absurdly detailed, automatically generated world model (with some power-ups manually added in). Why would a policy which manipulates the operator within the simulated world score well? It seems like it would take a confused world model for manipulation of the simulated operator to help with picking up simulated power-ups. Like if painting animals on cave walls actually caused them to appear. Larry Ellison is not going to win a yacht race by telling his data scientist to cripple his opponents in a simulation.

[Another frame: Cartesian dualism will happen by default, or at least will be easy to enforce on the architectural level. You could argue Cartesian dualists lose because they don′t do self-improvement? But an implied premise of your post is that foom won′t happen. I disagree but that′s another discussion.]

But let′s suppose the world model actually *is* confused, and the best policy in the simulation is one that manipulates the simulated operator to gain simulated power-ups. Even in this case, I think we′d still see a video like I linked earlier. We′d see the boat powering over to the part of the simulated world where the simulated operator resides, doing something to manipulate the simulated operator, and then the boat would have loads of power-ups somehow. I think the biggest concern is exposure to an information hazard when we see how the boat manipulates the operator. (Luckily, if we implement an information hazard filter before letting ourselves watch the video, the boat will not optimize to get past it.)

Human billionaires are hiring physicists to try & figure out if our universe is a simulation and if so, how to hack our way out. So there might be something here. Maybe if world model construction happens in tandem with exploring the space of policies, the boat will start ″considering the possibility that it′s in a simulation″ in a sense. (Will trying to manipulate the thing controlling the simulation be a policy that performs well in the simulation?)

[-] **Paul Christiano** 5y < 5 >

I′m not mostly worried about influence-seeking behavior emerging by ″specify a goal″ --> ″getting influence is the best way to achieve that goal.″ I′m mostly worried about influence-seeking behavior emerging within a system by virtue of selection within that process (and by randomness at the lowest level).

[-] **Alex Turner** 5y < 3 >

So the concern here is that *even if* the goal, say, robustly penalizes gaining influence, the agent still has internal selection pressures for seeking influence? And this might not be penalized by the outer criterion if the policy plays nice on-distribution?

[-] **Vladimir Mikulik** 4y < 2 >

The goal that the agent is selected to score well on is not necessarily the goal that the agent is itself pursuing. So, unless the agent’s internal goal matches the goal for which it’s selected, the agent might still seek influence because its internal goal permits that. I think this is in part what Paul means by “Avoiding end-to-end optimization may help prevent the emergence of influence-seeking behaviors (by improving human understanding of and hence control over the kind of reasoning that emerges)”

[-] **Alex Turner** 4y < 1 >

And if the internal goal doesn’t permit that? I’m trying to feel out which levels of meta are problematic in this situation.

[−] **Richard Ngo** 3y ‹ 2 ›

I recently came back to this post because I remembered it having examples of what influence-seeking agents might look like, and wanted to quote them. But now that I'm rereading in detail, it's all very vague. E.g.

> A few automated systems go off the rails in response to some local shock. As those systems go off the rails, the local shock is compounded into a larger disturbance; more and more automated systems move further from their training distribution and start failing.

This doesn't constrain my expectations about what the automated systems are doing in any way; nor does it distinguish between recoverable and irrecoverable shocks. Is AI control over militaries necessary for a correlated automation failure to be irrecoverable? Or control over basic infrastructure? How well do AIs need to cooperate with each other to prevent humans from targeting them individually?

Overall I'm downgrading my credence in this scenario.

Moderation Log