Leopold Aschenbrenner, August 10, 2021

Alignment report reviewer questions

Instructions for reviewers:

1. Please read the report and answer the questions below (short answers are fine, and no need to repeat content already discussed). When you are done, email a copy of this document with your answers to nick@openphilanthropy.org. We plan to share your answers with the author of the report (Joe Carlsmith) by default so that we can improve the next version of the report. If you want to pass anything on to me privately, please send it over email to nick@openphilanthropy.org.

2. If you would like to leave comments on the draft directly as well, feel free to make your own copy, label it with your name, and share it with nick@openphilanthropy.org when you are done.

3. Some questions involve assigning rough subjective probabilities to imprecisely operationalized claims.[1] Here, we're just looking for a loose sense of your epistemic relationship to the claim in question (we find that even rough probabilities in this respect are preferable to words like "plausible," "unlikely," "significant risk," and so on).

The author is happy to answer any questions you have over email, or on a call, while you are reading the report. Please feel free to reach out to him, if it would be helpful to you, at joseph@openphilanthropy.org.

[1] One way of defining these subjective probabilities is via preferences over lotteries. On such a definition, "I think it less than 10% probable that Kanye West will be the next president" means that if I had the option to win $10,000 if Kanye West is the next president, or the option to win $10,000 if a ten sided dice comes up "1", I would choose the latter option. See this blog post by Andrew Critch for more informal discussion; and see Muehlhauser (2017a), section 2, for discussion of some complexities involved in using these probabilities in practice.

General Feedback

Overall, I thought this report was extremely useful and enlightening. Thank you for this contribution! As someone less steeped in the AI debate, I am glad you are laying out the AI risk case in such a systematic way. It's been very helpful in clarifying my own thoughts on this issue.

Caveat up front: These were my views as of ~July 2021.

My two main disagreements with the report are about correction and catastrophe. My disagreements mean a substantially lower possibility of AI doom. But they also suggest to me that the AI doom scenarios really rely on quite specific premises, in particular either 1) long-run, correlated deception or 2) "foom"—rather than the more general case outlined in the report. For the future, I'd be most excited about work looking into the technical aspects of correlated deception and foom and determining how likely they are; results of such an investigation could substantially update my probability of AI doom.

Correction

- I think correction will be a lot easier than the report presumes. The APS system won't emerge in isolation, but in a multipolar world with lots of other advanced capabilities. If it starts wreaking havoc, we would likely be able to keep it in check.
- I see human society as a collection of depraved agents (individuals, groups, nations); they are kept in check by each other, by competition between them.
    - For example, I could tell you about an agent in the world that is very misaligned (downright evil, proto-fascist, depending on who you ask), has the brainpower equivalent to 1.4 billion people, is in control of the second largest military in the world, and is probably power-seeking.
        - This seems like a pretty worst-case version of the scenario outlined in the report.
    - This agent is called China!
    - Fwiw, I am quite worried about China. But I see the argument about AI as analogous to an argument about other misaligned agents. And my prior on other misaligned agents is that competition and correction is generally very good at keeping them in check.
    - For example, Amazon is not very aligned with overall human objectives, it has a ton of power, and yet our societal systems are able to keep Amazon in check, and more or less direct it towards overall human objections.
- So my prior on misaligned, agentic AI is that similar error correction mechanisms will keep them in check.

- A counterargument is that this APS system will be much smarter than everyone else, and so it'll be much easier for it to "take over the world" than for normal, human misaligned agents.
    - But arguments along these lines tend to implicitly presuppose a unipolar AI, something along the lines of AI "foom." Yes, if somebody in their garage can suddenly build advanced APS in a world that looks like our current world, we are doomed.

JC: Let's distinguish between:

1. A single PS-misaligned APS system will be much smarter and more capable than everyone else, so it will "take over the world."
2. Absent a competitive, scalable alignment solution, PS-misaligned APS AI systems will generally be much smarter and more capable than realistically-cooperating humans + their available controllable tools, to an extent that will put PS-misaligned AI systems in a position to gain and maintain the power they (by hypothesis) seek. This dynamic will lead to human disempowerment relative to some set of PS-misaligned AI systems, even if no single such system has "taken over the world," and even if such systems continue to compete amongst one another.

2 is the main concern in the report, and it doesn't require 1, unipolarity, or "foom."

Consider the example I discuss in section 6.3: that is, the relationship between chimpanzees and humans. No single human has taken over the world, and it's true that many humans/human institutions are misaligned with each other but hold each other in check by competing etc. But chimpanzees are still disempowered relative to humans as a whole. The worry is that humans as a whole will go the way of the chimps in this respect.

As a toy example, imagine multi-polar world dominated by very powerful, unconscious APS systems, each trying to e.g. maximize the number of clicks on a different website, but each at sufficiently similar levels of power that they can compete and hold each other in check – but where humans and human-aligned institutions are no longer competitive with any of them. Here I expect human disempowerment, and a click-centered outcome bad by human lights.

So "multi-polar world with advanced capabilities" isn't enough. We need an inference from "multi-polar world with advanced capabilities" to "realistically-coordinated humans + their controllable tools remain competitive." I expect a lot of our disagreement is about the strength of this inference. In particular, we can distinguish two cases:

A. Humans can create PS-aligned APS-AI, and doing so allows them to remain competitive with whatever PS-misaligned APS systems get deployed.
B. Humans are not able to create PS-aligned APS-AI, but they have sufficient advantages in other respects (e.g., from non-APS AI tools, practically-PS-misaligned APS systems that we succeed in getting to do useful work for us regardless, cybersecurity, compute

control, pre-existing military advantages, etc) that they remain competitive with whatever misaligned APS-AI systems get deployed.

My basic take is that (A) is great if you can get it, but it requires identifying methods of aligning your APS-systems that remain competitive with misaligned APS systems (e.g., that do not implicate too severe an "alignment tax"), even as frontier capabilities escalate. And the worry (see section 4) is that this might be too difficult.

(B), by contrast, mostly looks to me like a temporary fix if it works at all – at least assuming that frontier APS-systems continue to escalate in capability, and we don't successfully transition to (A) or otherwise achieve some serious and not-clearly-plausible degree of coordination in addressing the issue. My sense is that our main source of disagreement is about how much B gets you.

That said: I'll flag that I do, separately, think unipolarity a plausible outcome. This isn't because I expect something like "someone in their garage can suddenly build advanced APS-AI" – that feels to me like a straw man. Rather, I think it plausible, across a range of take-off parameters, that advanced APS-AI capabilities confer sufficiently serious and rapidly-escalating advantages in power-relevant domains (science, technology, strategy, persuasion, economic productivity, AI research itself, etc) that actors (including misaligned AI actors) with realistic leads are able to turn those leads into something like a Bostromian "decisive strategic advantage." I do think "why hasn't any human actor/institution taken over the world yet" is an objection this story needs to answer, but it seems answerable (salient to me, for example, are differences in the absolute level of tech power available, the pace/variance of tech change/growth in an advanced-AI regime, and projections from patterns of centralizing power over time, etc – see e.g. Bostrom's discussion in section 5 here).

- Realistically, the world in which these APS systems are deployed are worlds already replete with other AIs.
    - Human-level language models, lots of extremely capable narrow AIs, etc. These will give the humans opposed to the misaligned APS system ample "firepower" to continue error correction as always.

JC: I'm not sure where your confidence in the ampleness of this firepower comes from. To me it seems quite plausible that APS-systems will be systematically more useful for power-relevant tasks (e.g., science, technology, strategy, persuasion, AI research, economic productivity, etc) than narrow, non-APS AI systems and advanced language models, due to APS systems' ability to make and execute sophisticated plans, and their high-level strategic awareness of what's going on.

That said, I think the assumption that agentic planning and strategic awareness confer a competitive advantage in power-relevant domains (in addition to being more generically "useful") is a good one to bring out explicitly, and to question. It seems highly plausible to me,

but not guaranteed, and its falsehood would be a big source of comfort (its falsehood would also be relevant to the "Incentives" premise, in addition to the "Correction" premises).

(Also note that "generality" isn't strictly a part of the definition of an APS system. See discussion at the end of section 4.3.2.1.)

- For example, one pretty central premise to the APS system gaining a lot of power is hacking, so it can gain control over the physical environment (infrastructure, militaries). But here, I expect lots of narrow, very advanced "cyber defense" AIs.

JC: I do think that if humans have non-APS cyber-defense AIs that successfully prevent hacking by PS-misaligned APS-systems, this is a great help. But it seems plausible to me that APS-systems (or attackers more generally; cybersecurity today seems pretty non-great, though see Garfinkel and Dafoe for analysis of how things scale) will have an advantage here, too. And there are non-hacking ways of getting power, too (e.g., deceiving/persuading/coercing/trading with humans).

- While the alignment problem may be hard, I don't think it will be very hard to tell a system to "fight" another system (say, counter their online disinformation campaign, counter their hacking, defend against the opposing nation's roboarmy etc.)

JC: Again, I'm not sure where your confidence in the ease of telling one PS-misaligned system to "fight" another, in a way that you'll end up happy with, is coming from. By hypothesis, your PS-misaligned system is going to try to grab power on some inputs, so you need the ability to adequately control its opportunities and incentives to get the "fighting" behavior you want – a challenge closely continuous with achieving practical PS-alignment at all. And note that if you want your misaligned system to win the "fight," you'll be incentivized to empower it – e.g., give it more autonomy and resources, scale up its capabilities, etc.

- I find it implausible that all these systems (many different APS systems, most of all many different types of systems, like narrow AIs and language models) will somehow have some very correlated alignment failure that would prevent the competition between them.

JC: For different APS systems, to me there is a very salient and clear source of correlation in their degree of alignment: namely, whether we have figured out how to build PS-aligned APS systems at all. If we haven't, then it seems like everyone who is building APS systems will be building PS-misaligned ones.

And we can imagine correlation even if we've solved alignment at some level. Maybe, for example, we know how to align some APS systems, but only in a way that renders them uncompetitive (e.g., the solution only works on weaker systems, or via a very costly set of

additional training procedures and deployment constraints that create too large of an alignment tax).

There are also some broader reasons one might expect correlations in the properties of frontier systems. See e.g. Hubinger here.

- Importantly, lots of other advanced AIs in such a world would also help you align the APS systems (not just defend against mishaps/keep in check by competition).
    - E.g., a scenario I can imagine is that we get a particularly scary warning shot, but are still able to defend and correct and shut the rogue system down. That results in a moratorium on such systems, while we use the other capable, more narrow AI systems to work on an alignment solution.

JC: Early, less-scary systems helping us bootstrap to scalable alignment solutions is indeed one key way I think things could go OK. Counting on global moratoria on AI development seems like a big coordination ask, though – and the capability level "able to cause a warning shot sufficiently scary to prompt a coordinated global moratoria on an very broad, profitable and power-relevant industry, but not able to succeed in the relevant disempowerment attempt" seems like it might be a narrow band.

- I think particularly constraining here will be the compute-dependence of AI.
    - Whoever has the most compute will be more powerful. But compute seems uniquely well-suited to constraining by traditional, physical means.
    - Say somebody develops a misaligned APS system that starts wrecking havoc, taking over the world, etc. Well, send the US Navy to interdict all shipments of further compute! Send some Tomahawk missiles to destroy their data centers! Etc.
    - Some APS system taking over the world would require this APS system to somehow gain control over the majority of the world's compute, or at least a large fraction. Why would we let it? It seems quite easy to stop them.

JC: I think a compute-centric scenario is indeed helpful for tracking and responding to what's going on with the biggest/most powerful AI-related stuff, including misaligned stuff. That said, once trained (the biggest compute costs are in training rather than inference), I don't think a misaligned AI system needs a big fraction of the world's compute to pose a serious threat (building bioweapons, nano-technology, etc doesn't necessarily require a large share of world compute – though it does require infrastructure access). Also, as ever, I'm not assuming that a single AI system takes over (or that the worrying systems are confined to small sets of identifiable data centers, or that they are unable to defend/hide their key infrastructure, or to anticipate salient human responses to their strategies). And note that rogue AI systems can make software improvements that improve their compute efficiency; and that if you can't build aligned systems that would be competitive with the misaligned ones, your compute is less useful to you.

Beyond this, I worry you're assuming an implausible degree of easiness in controlling compute access in beneficial ways, in a world with lots of PS-misaligned APS systems running around amidst lots of uncoordinated humans. Misaligned AI systems can convince/coerce/deceive humans into giving them compute access; they can hack into or buy compute; worrying AI-related stuff might be happening under the auspices of human institutions and government; the candidate "controllers of compute access" might not be adequately confident that worrying that's happening; it might be too politically or militarily costly to interfere; they might not be convinced of the need for interference; there might be too many projects to interfere in this way in all of them, and so on.

- One compelling counterargument to all these lines of reasoning is: deception.
    - Suppose the APS system(s) seem (relatively) benign to us. And then because of competitive pressures, we all adopt such APS systems, gradually give it more and more control over a period of many years, give it control of our military, etc.
        - If then, 10 years later, the APS system comes out and says "Surprise! I was fooling you all along, and now I can take over the world because I am in control of everything bwahahaha!" then we are indeed screwed.
    - But this deception story is a to me somewhat implausible story: this system will be benign in all our training and extensive testing, it'll be benign throughout a period of many years as we gradually give it more and more control, it'll be so benign it'll convince all opposing foreign powers to adopt it (e.g. China adopting the US APS systems..), etc.
        - In particular, it needs to have full control of all the militaries. If there is a sufficiently large opposing military force, that could prevent the AI from taking over the world. This seems particularly implausible to me, that we will put the same or very similar APS systems in charge of all the militaries.

JC: I am not imagining a scenario in which a single deceptive AI system somehow convinces everyone to give it control over the whole world, over many years, while appearing totally benign.

I do think that if we can somehow eliminate deception from our APS systems, this is a huge help. But I also think that once a system is power-seeking and suitably non-myopic, various types of deception seem like they fall out of instrumental convergence arguments in a fairly straightforward way, so deception doesn't seem to me an especially exotic behavior to expect.

(Also, I'm actually not sure how much deception is strictly required for disempowerment. Consider, for example, the "superintelligent chimps" thought experiment in section 5.3.3: e.g., chimps that are able to be extremely useful to humans given the right incentives, but who are seeking food and entertainment for themselves. I can imagine some of these chimps deceiving humans in various ways, but even absent active deception, I can also imagine them just generally ending up with a lot of power, because they are so useful/profitable that some humans

are willing to use them despite recognizing that they'll seek power in misaligned ways given some inputs.)

- One story I could imagine is gradually giving control to not one system, but giving control to many different AIs over time.

JC: This is closer to the type of deception I'd have in mind, except that a lot of the power would be taken rather than given. (My sense is that you're assuming that all power-grabs short of world-take-over will be corrected, whereas I am not assuming this. For example, I think that if an AI system manages to hack its way out and copy itself onto the internet, from which point it starts engaging in various clandestine efforts to make money and enhance its capabilities, or if an AI convinces a corporation or set of humans to start doing its bidding, or whatever, then this kind of thing may well not get corrected, but rather become part of the background set of many problems that the world is dealing with.)

- But for human disempowerment, would need alignment of these AI systems to fail in a very correlated way, and all these AI systems to, from the beginning, collaborate to collectively deceive us.
    - I see the question of "will all the AIs be misaligned in a very correlated way and deceive us collaboratively" as fairly central.
    - It seems particularly hard for many different AIs to collectively deceive us. They'd all have to fail in similar way right from the start. There would be many opportunities for warning shots and defection.
    - As I mention above, it seems particularly straightforward to tell an AI to "fight" another AI (e.g. US and Chinese military AIs will be programmed to fight each other/defend against the other). It would have to be a particularly extreme failure to overcome this basic programming and for these AIs to instead collaborate.
    - Note there will be very strong incentives for, say, the US military AI to not fail to fight/defend against the Chinese military AI.

JC: I think it's important to distinguish between (a) correlated PS-misalignment, (b) correlated deception, and (c) active collaboration amongst PS-misaligned systems (whether on deception, or in general). I see correlated PS-misalignment as very plausible, where the source of the correlation is the degree of progress the world as a whole has made on PS-alignment. Conditional on lots of misalignment, I do expect lots of deception, but not in some systematic way coordinated amongst all the AI systems "from the beginning." And I'm not assuming collaboration between many different AI systems (though I do think this is possible).

It seems like you think collaboration is necessary centrally because otherwise you get warning shots and defections, plus competition more generally. Warning shots aren't all that much comfort to me, though (see section 6.2). Defections (e.g., one AI system truthfully exposing another's deception/misalignment) do seem helpful if you can adequately incentivize them (and I think we should try to do this), but I lump it in same bucket of "temporary fix, if it works at all"

that I put most efforts to try to constrain PS-misaligned APS-AI systems without actually solving the alignment problem in a scalable way. And as discussed above, I think competition between PS-misaligned AI systems only useful if it renders *humans* competitive – and I think I'm more pessimistic than you about our ability to make use of PS-misaligned APS-systems in ways we end up happy with (including – maybe especially? – in military contexts).

That said, in multi-polar and slower-take-off scenarios, I do broadly expect pretty complicated collaboration/trade/bargaining dynamics between humans and not-yet-crazy-powerful misaligned AIs, rather than a single "AI bloc" united against the humans – and I do think this is a source of comfort.

- All of the points above don't necessarily reduce the probability of a climate change-scale catastrophe. Rather, my claim is that if an AI starts causing damages (on the order of climate change), we'll very likely be able to stop it from fully Terminator-style taking over the world.
- One more general point I would add here: if I look at the world today and how we approach new technology and growth (nuclear energy, biotech, the FDA on vaccines, NIMBYism, etc.), if anything I think we are wayyy too risk-averse with respect to new technology than too risk-loving.
    - So this informs my prior that society will generally be very risk-averse and safety-oriented with respect to these AI systems—and be willing to take extraordinary steps to shut them down if they start going awry.
    - We shut down the J&J vaccine for astronomically small side effects, while the British covid variant was rampaging the country!
    - Similarly, no country tried human challenge trials in 2020! Even though this could have allowed them to deploy the vaccine 6+ months earlier, and this country would have gained a huge competitive advantage!! No country even tried selling the vaccine on the free market (i.e. not banning it prior to ~Dec 2020). All the arguments you could make about AI and competitive pressures apply to this example, and yet we were still way too risk averse, not too risk-loving.

JC: The fact that humans are often risk-averse with new tech does seem helpful to keep in mind (though there are some other examples, like gain-of-function research, that look like they go in the other direction).

Catastrophe

- I think the probability of existential catastrophe, conditional on disempowerment, is a lot lower than the 95% in the report, for two reasons.
- First, a conditional probability point. In the worlds where we let an AI take over and fail to stop it, I think it's more likely to be only subtly misaligned.
    - See discussion above. In scenarios in which AI is very misaligned, with fundamentally bad objectives, I think it's likely we'd try very hard to stop it and likely succeed.

- So most of the probability mass where we let AI take over is where the AI is only subtly misaligned, such that we are pretty blase about it taking over.
- For example, maybe the AI has a goal along the lines of "maximize GDP." It understands what GDP is (so it doesn't do something dumb like hack the GDP statistics). GDP is pretty good proxy for human welfare, so the end result is pretty good. But there are some flaws with GDP, so maybe the AI suboptimally neglects e.g. opera (AI doesn't sufficiently take into account the intrinsic value of culture).
- An example you often give for why we would let AI doom happen is the example of climate change. But that's exactly my point! Sure, catastrophes and subtle misalignment on the (small) scale of climate change might well happen. But that's very different from Terminator-style scenarios. When the Nazis threatened to take over the world, we got our act together and stopped them; the Nazis are a lot worse than climate change.
- A fundamental point here is that I don't see current human society as all that "aligned" anyway. For example, I think we currently do something along the lines of the "maximize GDP" example above, so AI doing that would be no worse. In fact, I think an AI solely trying to "maximize GDP" might actually do a lot better than current human society (e.g. making it more likely that we colonize space, vs. being sedentary and decadent and stuck on this planet). So I am pretty unconcerned with subtly misaligned AIs taking over.
- TIdr; I think the modal outcome of AI disempowerment in this conditional probability scenario is more like your disfavored political party taking over (say, the Liberal Democrats rather than Labour), rather than future Nazis taking over the world, or something even worse.
- This point is probably mostly a problem I have with the framing of "disempowerment."

I disagree with this along a number of dimensions. First, I don't think that conditional on the first five premises, the PS-misaligned systems that have disempowered humanity are likely to be "only subtly misaligned." It sounds like you're thinking that the main way humans as a whole get unintentionally disempowered is via their not having a sufficiently strong preference for remaining in power, because correction is sufficiently easy that (absent foom or long-term correlated deception), humans would just choose to remain empowered if they really wanted to. As indicated above, I disagree that correction is this easy: I think that in most cases where 1-5 hold, humans really wanted (or would've wanted) the process of disempowerment to not be happening, but it happened anyway because they couldn't stop it.

Beyond disagreements about ease of correction, though, I am also more skeptical that the notion of "only subtly PS-misaligned systems, leading to not-that-bad futures" is likely to apply in many realistic worlds. Part of this is about the idea of "value fragility" -- e.g., that the space of values that you're happy to see optimized in extremis is actually quite small, and closely related to contingencies of human-ness (though I do think there are important questions to ask about value fragility, some of which your comments re: the alignment of humans with each other point at -- it's a topic I'm hoping to think more about). For example, I'm not sure exactly what your

imagined “humans are all involuntarily disempowered, but the PS-misaligned AI systems that disempowered them are maximizing ‘GDP-as-truly-understood’ (not just some metric of GDP)” world looks like, but my guess is that I don’t like it very much, and that per the “problems with proxies” discussed in section 4.3.1.1, the maximization in question breaks current correlations between GDP and human welfare. And note that the type of bad that I’m worried AI values will be is not “fundamentally evil” in the sense your Nazi example calls to mind. Rather, it’s just “orthogonal to what we ultimately care about” -- and the worry is that (conditional on sufficient optimization) such orthogonality is the strong default.

It seems like part of what’s animating your view here is a certain kind of pessimism about the default trajectory of human civilization absent AI catastrophe, and about the values of most present-day humans -- pessimism that makes turning the future over to misaligned AI systems seem like less of a blow to its value. I expect that I am more optimistic than you on this front, especially relative to random AI objectives like maximizing clicks on websites, but I grant that sufficient pessimism of this kind can alter one’s sense of the stakes, here.

- Second, it seems quite plausible to me that the AI will have better “values” anyway.
    - Humans were programmed by evolution (a very naive, dumb optimization algorithm) and yet with have ended up with a sophisticated morality. Why won’t the AI do the same?
    - If anything, a smarter AI seems to me would be more likely to know what is “right” than we do.
    - As a prior, I don’t know if we can distinguish the AI from our evolutionary situation, so as a prior, I would take a 50% haircut on P(catastrophe) from P(disempowerment).

I’m unsympathetic to the idea that PS-misaligned AI systems will end up pursuing actively *better* values than humans given access to comparable cognitive capabilities (e.g., aligned AI systems helping them) would, by default. As I briefly touch in on section 7, I don’t expect all cognitive systems to converge on the “right values” given sufficient understanding (see e.g. “the orthogonality thesis”), and my own meta-ethical view is that the “right values” for us to pursue are closely related to contingencies of what we in particular happen to care about -- contingencies that I don’t expect to apply to PS-misaligned AI systems by default (even ones produced via naive, dumb optimization algorithms).

Maybe lots of intelligent systems converge on certain types of cooperative moral-ish behavior for complicated instrumental reasons, but my best guess (I’m actually writing a blog post on the topic at the moment) is that this doesn’t go as far as agents lacking in relative power would like. Regardless, though, I don’t expect similarities between gradient descent and evolution to play much of a role in ensuring the relevant forms of cooperation. Do you think that a system trained via gradient descent to maximize clicks on a website will end up with something resembling human morality (or something better), in virtue of its training process rather than its (eventual) intelligence? I do not. And if such a system doesn’t end up with a human-like morality, but is instead mostly trying to maximize clicks, are you agnostic about whether it’s values, or yours,

are more worth pursuing? If so, I expect there is some deeper meta-ethical disagreement here, and one that I expect to imply a kind of wholesale moral skepticism on your end (see e.g. Street's Darwinian Dilemma for an example of the type of dynamic I have in mind).

As a final general point, I think it would be worth quantifying how much of the AI risk comes from "feasible alignment challenges that we could solve with a few extra decades of more time doing technical work" vs. "black swan tail risk that we can't do much about anyway." A lot of the arguments I hear about decade-long correlated deception feel more like the latter, though I am unsure. It would be extremely decision-relevant to know how much of the AI xrisk we can actually affect. If the xrisk is mainly of the latter, inevitable type, then there's not much we can do anyway, and so our decisions should be focused on reducing other risks (e.g. risk from China getting AI first).

I agree that how much we can influence the level of risk here is important -- but that question is outside the scope of the present project, the aim of which is just to assess the absolute risk level.

Overall

1. *This report aims to (a) articulate and evaluate an overall argument in support of the conclusion that misaligned, power-seeking AI systems will cause an existential catastrophe before 2070, and (b) to assign rough probabilities to the premises (and ultimately, the conclusion) of this argument.*

   Is the report's main project framed in a way that makes sense to you? If not, what is unclear or confusing?

   Yes. I think this is a (the?) central decision-relevant question for OpenPhil.

2. *Please briefly summarize the key argument of the report and, where applicable, which pieces of the reasoning you found most interesting/informative.*

   Overall argument is clear, and I thought it was generally well laid out. There were many neat pieces of reasoning, and many times an objection came up in my mind and you responded a few paragraphs later, which I appreciated.

3. *Did you find the report overall clear, reasonable, and convincing? If not, what seemed unclear, unreasonable, or unconvincing?*

   Yes; re unconvincing see general comments above.

Timelines

The report focuses on AI systems with three key properties (the author calls these “APS systems”):

- *Advanced capability*: they outperform the best humans on some set of tasks which when performed at advanced levels grant significant power in today’s world (tasks like scientific research, business/military/political strategy, engineering, and persuasion/manipulation).
- *Agentic planning*: they make and execute plans, in pursuit of objectives, on the basis of models of the world.
- *Strategic awareness*: the models they use in making plans represent with reasonable accuracy the causal upshot of gaining and maintaining power over humans and the real-world environment.

4. *Any comments about or objections to this framing?*

   I think a lot (most?) of the action is happening in “strategic awareness.” The capability for long-range planning of this sort seems both 1) very hard, much harder than e.g. human-level language models or other very capable AIs, and 2) key to arguments about likelihood of deception. I would have appreciated a more detailed treatment here.
   I also thought the difference between agentic planning and strategic awareness was a bit fuzzy to me.

   In particular, I think the long-range planning needs to be extremely advanced for long-range deception to be plausible. To pull off deception, the AI needs to go through extensive testing and appear benign, very extensive deployment in the real world and appear benign, and coordinate with other similar AIs, each of whom need to do the long-range planning and appear benign, to gain enough power before “revealing its true intentions.”

JC: I agree that the long-range planning in question here probably needs to be pretty sophisticated (though I have more than deception in mind, and as noted above, I don’t think coordination necessary), and that this degree of sophistication isn’t directly implied by e.g. human level language models. But I’m explicitly forecasting something more advanced than human-level language models.

Re the difference between agentic planning and strategic awareness: agentic planning is about using models of the world to pursue objectives, strategic awareness is about the sophistication of those models.

5. *What rough probability would you place on the following claim?*

TIMELINES: By 2070, it will become possible and financially feasible to build APS systems.

> The 65% in the report seems too high. I would more generally put a substantially lower probability on APS systems like this than very powerful/human-level AI in general.
>
> I might put 25% on this.
>
> I think this sequencing (APS happening later than otherwise extremely capable AIs) matters, e.g. for difficulty of alignment and feasibility of correction. However, I am not an expert in this; I would have appreciated more of a discussion of what you expect the AI landscape to look like pre-APS.

Incentives

The author assumes that there are strong incentives to automate advanced capabilities, and discusses three reasons we might expect incentives to push relevant actors to build agentic planning and strategically aware systems in particular, once doing is possible and financially feasible:

- Usefulness. Agentic planning and strategic awareness both seem very *useful*. That is, many of the tasks we want AI systems to perform seem to require or substantially benefit from these abilities.
- Available techniques. Given available techniques, it may be that the most efficient way to *develop* AI systems that perform various valuable tasks involves developing strategically aware, agentic planners, even if other options are in principle available.
- Byproducts of sophistication. It might be difficult to *prevent* agentic planning and strategic awareness from developing in suitably sophisticated and efficient systems.

6. *Do you find these reasons persuasive? If not, why not?*

> Yes, with the most weight on usefulness.
>
> One way I thought about it: it seems quite plausible to me that AI-driven explosive growth will *require* APS systems of this type. “Mere” oracle AIs and similar systems might (are likely?) to still be complementary, rather than substitutable, to human labor. If it’s true APS is required for explosive growth, then there would seem to be extremely large incentives here.

7. *What rough probability would you place on the following claim?*

INCENTIVES: By 2070, and conditional on TIMELINES above, there will be strong incentives to build APS systems.[2]

> Joe gives 80%; that seems right to me.

Alignment

In this section, author discusses the hypothesis that for APS systems, misaligned behavior on some inputs (where this behavior involves agentic, strategically-aware planning in pursuit of problematic objectives) strongly suggests misaligned *power-seeking* on some inputs, too (in brief, the central argument here is that power is useful for pursuing a wide variety of objectives).[3] He then discusses different challenges to ensuring that APS systems are practically PS-aligned. In particular, he considers two key challenges to controlling a system's objectives adequately:

- *Problems with proxies*: Optimizing for a proxy correlated with intended behavior may break the correlation.
- *Problems with search*: Searching over systems that meet some evaluation criteria won't necessarily result in systems intrinsically motivated by those criteria.

He also discusses the possible role of:

- restricting the temporal horizons of an APS system's objectives;
- controlling its capabilities (for example, by giving it a specialized capability profile, and preventing problematic capability improvements);
- controlling its options and incentives (for example, via restricting the environments it operates in, monitoring its behavior, and providing incentives towards cooperation).

Finally, he discusses three ways that ensuring the safety of APS systems seems unusually difficult, relative to safety challenges posed by other technologies. Namely:

---

[2] Here we understand "incentives" in a manner such that, if people will buy tables, and the only (or the most efficient) tables you can build are flammable, then there are incentives to build flammable tables, even if people would buy/prefer fire-resistant ones.

[3] Definitions:

- *Misaligned behavior*: unintended behavior that arises in virtue of problems with a system's objectives.
- *Misaligned power-seeking*: misaligned behavior that involves seeking to gain/maintain power in unintended ways.
- *Practically PS-aligned*: a system doesn't engage in misaligned power-seeking on any of the inputs it's in fact exposed to.

- The behavior of agentic, strategically aware systems much more cognitively sophisticated than humans may be uniquely difficult to predict and understand. This issue is especially salient in the context of contemporary machine learning, in which our ability to create an AI system that can perform some task (e.g., predicting text) often far exceeds our ability to understand *how* the system does what it does.
- APS systems may be actively and adversarially optimizing for getting deployed, including via deceiving and manipulating us. This makes it harder to trust things like safety tests.
- The stakes of error are unusually high (e.g., bioweapons may be a better analogy than planes or bridges).

The author concludes that ensuring the practical PS-alignment of APS systems could well be difficult.

8. *Do you find the author's discussion and conclusion in this section persuasive? If not, why not?*

   Generally, yes.

   However, I thought it was a bit too focused on general alignment issues vs. alignment issues that are specifically ones that could lead to xrisk (in particular, again, deception). I have a strong expectation that more basic alignment issues (dealing with proxies, generally doing roughly what you want it to do in the immediate future, etc.) will be solved anyway for deployment because they are necessary for the AIs to be useful at all.

   I also think it would be worth considering more the environment in which these APS systems will likely emerge—replete with other AI systems—and how this will make alignment easier. E.g., can we use superintelligent oracle systems to help with aligning APS? That sort of thing seems very plausible to me. I think my longer timelines for APS systems (vs. otherwise capable AI systems) are an important part of my optimism here. I think the truly dangerous APS systems will come quite a bit later than simpler AI systems; we will have a lot more experience and tools then. Alignment is hard, but so is building the APS systems in the first place.

JC: I do think that aligned, superintelligent oracle systems, if you can get them, would help a lot: the question is whether you can actually build such systems in a way such that (a) they actually give fully truthful and honest answers, (b) they don't end up pursuing more agentic, long-term goals (see e.g. section 3.3), and (c) they can be used to create scalable alignment solutions for APS-systems, or to otherwise prevent problems from PS-misaligned APS-systems from arising even as an increasing number of relevant actors are in a position to create such systems.

Deployment

This section discusses why we might expect to see practically PS-misaligned systems actually get deployed. The author briefly discusses the possibility of unintentional deployment, then lays out various factors that might affect the beliefs and incentives of relevant actors in choosing to deploy a system that is in fact practically PS-misaligned (even if they don't know it). He focuses on four factors that seem especially concerning:

- Externalities and competition. The profit/power at stake in deployment may make it individually rational for actors to take risks that society as a whole (let alone all future generations) would not accept; and competition between actors may incentivize risk-taking.
- Over time, larger and larger numbers of relevant actors, with varying degrees of caution and social responsibility, may be in position to build these systems and take these risks.
- Practically PS-misaligned systems may still be able to demonstrate a lot of usefulness.
- Practically PS-misaligned systems might deceive or manipulate relevant decision-makers.

Overall, the author finds it plausible that if ensuring practical PS-alignment proves challenging, practically PS-misaligned systems could end up getting deployed.

9. *Do you find the discussion and conclusion in this section persuasive? Why or why not?*

> I think my points in the general comments are relevant here: 1) the types of PS-misaligned systems that will still get deployed are going to be the ones that are only subtly misaligned, on the scale of the climate change example repeatedly given; 2) I really do think deception is key. and I would have appreciated much more extensive discussion here.

See my response above.

10. *What rough probability would you place on the following claim?*

ALIGNMENT DIFFICULTY: By 2070, and conditional on TIMELINES and INCENTIVES above, it will be much harder to develop APS systems that would be practically PS-aligned if deployed, than to develop APS systems that would be practically PS-misaligned if deployed (even if relevant decision-makers don't know this), but which are at least superficially attractive to deploy anyway.

> Starting here, I find the framing of the probability steps less useful, in particular because of the difference between subtly misaligned vs. catastrophically misaligned.
>
> I'll accept Joe's 40% here, but for me the vast majority of that probability mass is "subtly misaligned" rather than "crazy 20 year deception scheme misaligned."

JC: See response above re: why I don't see "subtly misaligned" and "crazy 20 year deception scheme" as the main categories of misalignment to consider.

Correction

This section discusses whether or not we should expect the impact of deploying practically PS-misaligned APS systems to scale to the permanent disempowerment of ~all humans. It also discusses a few mechanisms relevant to the plausibility of this disempowerment. The author suggests that:

- Serious risks can arise even if frontier capabilities do not escalate very rapidly/discontinuously, or via an "intelligence explosion," or in a manner that concentrates power in the hands of a single actor.
- Early warnings/indications of problems are helpful, but not necessarily sufficient to solve the problems in question.
- Corrective feedback loops could be undermined by faster/more discontinuous escalations in capabilities, adversarial/deceptive dynamics, and coordination difficulties.

Overall, the author thinks that humans might well be able to correct PS-misalignment problems and prevent them from re-arising, but that doing so will likely require addressing one or more of the basic factors that gave rise to the issue in the first place: e.g., the difficulty of ensuring the practical PS-alignment of APS systems (especially in scalably competitive ways), the strong incentives to use/deploy such systems even if doing so risks practical PS-alignment failure, and the multiplicity of actors in a position to take such risks. This task, he thinks, could well prove difficult.

11. *Do you find the discussion and conclusion in this section persuasive? Why or why not?*

    See general comments.

12. *What rough probability would you place on the following claim?*

HIGH-IMPACT FAILURES: By 2070, and conditional on TIMELINES, INCENTIVES, and ALIGNMENT DIFFICULTY above, some deployed APS systems will be exposed to inputs where they seek power in misaligned and high-impact ways (say, collectively causing >$1 trillion 2021-dollars of damage).

    I'll accept Joe's 65% probability here.

    Climate-change-scale damages seem quite plausible to me. A committed terrorist group could already do a lot of damage, and they don't have powerful AIs yet!

13. *What rough probability would you place on the following claim?*

DISEMPOWERMENT: By 2070, and conditional on TIMELINES, INCENTIVES, ALIGNMENT DIFFICULTY, and HIGH-IMPACT FAILURES above, some of this misaligned power-seeking will scale (in aggregate) to the point of permanently disempowering ~all of humanity.

> 20%? (vs. Joe's 40%)
>
> See general comments; I think a thrust of my disagreement is here.
>
> Here, I'm focusing on/mostly counting only disempowerment that is catastrophic, vs. subtle misalignment.

Catastrophe

The author briefly discusses the possibility that the permanent and unintentional disempowerment ~all humans would not constitute an existential catastrophe (e.g., an event that drastically reduces the value of the trajectories along which earth-originating civilization could develop). He also emphasizes that some AI systems might have moral status, and that the right way forward may ultimately involve humans intentionally ceding power to them: the point in the present context is to avoid *unintentional* disempowerment of humans.

14. Any comments on this section?

> See general comments.

15. What rough probability would you place on the following claim?

CATASTROPHE: By 2070, and conditional on TIMELINES, INCENTIVES, ALIGNMENT DIFFICULTY, HIGH-IMPACT FAILURES, and DISEMPOWERMENT above, this disempowerment would constitute an existential catastrophe?

> 50%? (vs. Joe's 95%)
>
> Here my different views start jibing less with Joe's probability framework. (I.e., if I counted more subtle disempowerment in the probability step above, I would put a lower probability here.)

Overall probabilities

The author concludes by listing his own subjective, highly-unstable probabilities on each of these premises, along with a number of caveats about these probabilities should be understood. The probabilities are:

> *By 2070:*

1. It will become possible and financially feasible to build APS systems. **65%**
2. There will be strong incentives to build APS systems | (1). **80%**
3. It will be much harder to develop APS systems that would be practically PS-aligned if deployed, than to develop APS systems that would be practically PS-misaligned if deployed (even if relevant decision-makers don't know this), but which are at least superficially attractive to deploy anyway | (1)-(2). **40%**
4. Some deployed APS systems will be exposed to inputs where they seek power in misaligned and high-impact ways (say, collectively causing >$1 trillion 2021-dollars of damage) | (1)-(3). **65%**
5. Some of this misaligned power-seeking will scale (in aggregate) to the point of permanently disempowering ~all of humanity | (1)-(4). **40%**
6. PS-misaligned systems permanently disempowering ~all of humanity will constitute an existential catastrophe | (1)-(5). **95%**

In combination, these probabilities yield an **overall estimate of ~5% chance of an existential catastrophe by 2070 from scenarios where all of 1-6 are true**, which the author would adjust upwards to reflect power-seeking scenarios that don't fit some of 1-6. The author also notes in a footnote that his "high-end" and "low-end" estimates vary considerably: from between ~40% on the high-end, to ~.1% on the low end.

16. *Any comments on these probabilities? Does anything stand out to you as unreasonable?*

    See earlier comments, in particular on finding the last few probability steps not as useful because of subtle vs. catastrophic misalignment.

17. *What is your own rough overall probability of existential catastrophe by 2070 from scenarios where all of 1-6 are true?*

    The probabilities I gave differ in the following ways:
    - 25% instead of 65% on APS possibly by 2070
    - 20% vs. 40% on misaligned power-seeking will scale to disempowerment of humanity
    - 50% vs. 95% on catastrophe.

    Conditional on Joe's timelines, I end up with ~1.3%.

    With my timelines, I end up with ~0.5%.

18. *Any final comments on the report as a whole?*

    Great work! Very excited that you are working on this, and I'm very much looking forward to seeing any future work you do. I'd be particularly interested in a more extensive discussion of the technical likelihood of correlated, long-range deception.

    Thanks!

Permissions

19. *Would you be OK with us making your answers to the above questions publicly available?*

    Yes.

20. *Would you be OK with us publishing your name alongside your answers? Publishing with your name is preferable on our end, because we think it helps give readers more context, but it's also fine if you prefer to remain anonymous.*

    Yes.