Notes on Joe Carlsmith's AI Alignment Series

Brian McPeak

August 20, 2026

Over the last year or so, Joe Carlsmith has been releasing a series of essays on AI alignment. I read these and wrote down my thoughts in a post on my Substack; this document includes a more extended summary and thoughts on each essay. Each section will be about one essay, and the heading will be the essay's title.

1 What is it to solve the alignment problem?

The first essay is titled "What is it to solve the alignment problem?" and the goal is to answer that question. The short answer is that "solving the problem" means that we build super-intelligent agents and induce them to do world-improving work on our behalf, without suffering a *loss of control scenario*.

A loss of control scenario, or LOCS hereon, is when rogue AI actors have effectively taken over the planet. Rogue actors are those which engage in power-seeking behaviors despite not being designed to do so. A list is given of such behaviors, but they basically boil down to being some mix of harmful, untrustworthy, and disobedient, often in an attempt to gain more power. The idea that intelligent agents will seek power regardless of their goals is called *instrumental convergence* and is one of the load-bearing ideas in alignment philosophy.

LOCS are obvious in practice (you'll know it when you see it). The term does not refer to situations where control is willingly given to the AI. Indeed, the amount of control given to the system before it exhibits rogue behavior is an important parameter characterizing the scenario: realities where the ASI talks the researcher into letting it out of the box (*Ex Machina*), and then wreaks havoc, are very different from the scenario where the initially benevolent AI controlling all critical functions of society starts drifting into rogue behavior (*I, Robot*).

There are lots of ways that ASI might benefit us through breakthroughs in science and technology. The essay is not particularly interested in categorizing them. It's worth pointing out explicitly that the essay is interested in AI *agents*. There are other ways of getting the benefits of superintelligence. Perhaps we will make non-agentic highly specialized AI systems, or we will find a way to enhance our own biological intelligences.

An important subclass of AI benefits are *transition benefits*, which are those benefits that pertain to readying the world for the AI transition. Perhaps importantly, for the question of actual AI strategy, some of these benefits might be reaped with AIs that are not ASIs.

"Failing at alignment" means failing at safety—suffering a LOCS. There are different ways not to do this: you can *avoid* the problem by never building superintelligent agents; you can *handle* the problem by building them but failing to get them to use their full powers (maybe by only using them

on a small subset of problems). It also may be possible to still gain some or all of the benefits of superintelligence without agents. Any case of gaining the main benefits without a LOCS is victory, i.e. solving alignment.

> I'll make a note here. He says that he will focus on alignment because there's value in trying to attack an especially hard version of the problem. But does this imply that avoiding / handling alignment is less hard? For instance, it's plausible that avoiding the problem is impossible in a world with so many actors: someone will eventually build superintelligent agents. I think it's fair for him to focus on alignment, but perhaps these others are underexplored. "How could we, in practice, actually avoid / handle the problem?"

There are stronger definitions one might take of "solving alignment." *Safety at all scales* would require that we can safely access the benefits of even maximally superintelligent agents. An even stronger condition is *fully competitive safety*, which means that you can do this and remain competitive with even incautious actors. This might be the case if solving alignment is actually helpful to building effective AIs. Another goal might be to be *permanently* safe from LOCSs. And a weaker definition is to remain safe in the near term, however defined.

An orthogonal goal would be to optimize for what is sometimes called *coherent extrapolated volition*, which means that the AI does what humans want—or more specifically, what we would want if we were as smart and knowledgeable as the AI and if we had perfect self-knowledge and self-control.

Reasons are given for why each of these is not considered. Let me summarize the goals:

- **Safety at all scales:** it's too difficult to try to solve that problem now, and we will probably need help from intermediate AIs. So let's try to make sure that those are safe.
- **Fully competitive safety:** reason not explicitly given, though he hints that perhaps even incautious actors will become cautious as they better understand the risks, or their AIs resist building further unsafe generations.
- **Permanent safety:** would risk exerting permanent control over the future—perhaps this would stunt or stem human growth and flourishing?
- **Near-term safety:** attempting to limit the scope of the series. His definition brings out the technical aspects of alignment into clearer focus.
- **Complete alignment:** the reason to do this requires a sort of entirely different view on the alignment problem (e.g. value fragility) that he doesn't necessarily want to grant. Plus it's a very high bar, and we can probably do great good without solving it.

I think that from these reasons, I can see a glimpse of where this is going. The strategy will be to use AI to align AI. Namely, we don't aim at shifting ourselves into a basin where no loss of control scenario could ever occur. Rather, we figure out a strategy for getting AI to help us extract as much benefit as possible from AI, assuming finite timelines.

The author explains nicely his vision for the point of all this: we want to work towards good things. This will almost certainly involve us (humans) changing our role as AIs come to do more and more, and perhaps we might even be better off delegating everything to them. And ultimately, the reason to be most interested in navigating the short term is that some of the questions after that belong

to future generations. The first job is just to set them up as well as we can. I find this impulse very admirable.

I'll end the summary with an interesting question posed in Appendix 2. Could there be a case where a LOCS is actually *good*? This might be true if, for instance, AIs have the capacity to suffer, and do. The author even worries that his work might be evil, as it might be used to control conscious beings. This is obviously a huge separate question.

Questions and comments

- This is a bit of a nitpick, but it's not totally clear to me how much "solving alignment" requires that we understand what we're doing. It seems from the definitions that we win on safety by merely making it out alive. But one might instead want to make it out alive in many realistic / plausible / similar worlds, or that we make it out alive *because of* our scientific understanding of alignment.

- Why the focus on agents? Is it clear that agents are actually different from other forms of superintelligence?

- RE: is it wrong to try to control the AIs... it's murkier than that, right? Suppose that AIs can be conscious; then part of alignment is choosing which conscious beings to bring into the world. In some ways, *not* doing alignment is an evil injustice to all of the good, happy AI minds that might exist if we don't figure out how to prevent the (potentially non-conscious) paper-clip maximizer from flattening the world. In any case I agree that it comes down to bets and tradeoffs. More will need to be said.

- Isn't there a large batch of problems where generation alpha willingly gives all the power to the AI, and generations beta and on are left to deal with it? I know that only a finite number of things can be discussed, but I think some consideration of this type of problem—or more generally, the problems with handing over power and not being able to get it back—would be interesting to hear more about.

2 When should we worry about AI power seeking?

What are the conditions under which an AI might seek power? First, the AI should be sufficiently agentic that it is able to model the world and act on it in ways that allow and incentivize having power. AIs which don't have the ability to make plans, or which don't recognize that power would help them, won't seek power. Second, the AI's motivational structure must support gaining power as an instrumental goal. And finally, they must be incentivized to gain power—it must be a rational means to their ends.

To satisfy the agency prerequisite for power seeking, an agent must be capable of searching over plans. And to some extent, they need to act on the plans that they generate. They must be able to choose and act on specific plans, and their choices must be coherent enough across time that they actually fulfill at least some plans (so, for instance, half-way carrying out many different plans will not lead to success—see my research career for an example of this). And finally, their search needs to include rogue options. If we somehow constrain the search space to only good options, then the AI is not agentic enough to go rogue.

The motivational aspects are simply that the AI cares about goals which can be furthered, or

outcomes which can be made more likely, through its behavior. Furthermore, these motivations need to apply on sufficiently long timescales that it would have time and reason to gain power. A ChatGPT that only cares about giving you the best answer to the current question might have little reason to gain power. One which worries about being in a position to give the best answers to all questions for all time might be motivated to... explore outside-the-box solutions.

As for incentives, there are several distinctions to be made. First, note that the choice can be narrowed down to the AI's best non-rogue action and best rogue action. The choice will then be made using at least these four factors: how good the best non-rogue action is; how good (the result of) the best rogue action is; how much it disfavors the actions needed to obtain that rogue result; and how bad / likely the world is where it gets caught / stopped / fails to gain power.

From this framework, it is easy to see that there are different ways to incentivize the agent to choose non-rogue actions. We can make the non-rogue actions better, per the agent's estimations. We can make the payout of the rogue actions worse, we can instill "better morals" so it more strongly disfavors the rogue behavior, or we can make the chance and penalty of failure higher.

Indeed, there is a spectrum of solutions here. At the extremes, the AI has no rogue options so its motivations are irrelevant (re: safety), and the other pole where it can be trusted with arbitrary amounts of power, so its options don't matter. The author makes the case that the middle ground has been neglected.

It seems clear, especially today, that AIs will and do satisfy the agentic planning criteria. It's not necessarily true that they do or will be able to consider the bad options, but I agree with the author that they probably will. It just seems unlikely that such a handicapped system could be as capable as a freer one. The time-horizons criterion is less clear, but my guess is that they will pass it, especially as their ability to act quickly exceeds ours. Only the dimensionless number matters.

It's less clear if future agents will satisfy the incentive criteria. The author points out that scenarios more focused on in the alignment literature so far often involve a single AI actor who can take over the world easily. In this case, the incentive calculations simplify—we don't need to consider their fear of failure, only their intrinsic inhibitions for the actions they would need to take themselves.

Will the first superintelligence have a decisive strategic advantage? We don't know, but the world will already be populated by many other AIs. It is still plausible that there will be some sort of explosion in intelligence, but recent progress has been steady, with steeply increasing costs per unit increment in abilities.

Let us now define a scenario where humanity's continued empowerment relies on our ability to restrict the options of superintelligent agents, and their choosing not to (in other words, this essentially means that, based on the AI's capabilities, such options would exist). In such a case we might be said to be *globally vulnerable*. To be more precise, a vulnerability is a case where, absent option control (or in the nebulously defined "default scenario"), a rogue option would exist. So either option control or motivational control is necessary to prevent rogue action. A single AI with decisive strategic advantage would be one such case, but other weaker forms exist. Examples given include a superintelligent AI who would have a DSA if your attempts at option control fail, and cases where many dangerous but non-superintelligent AIs might coordinate to take power or behave roguishly. Importantly, the author says that his view is that global vulnerability is the default path. We will probably need some form of successful option or motivation control! Because given current trends, it seems likely that we are on course for highly capable AI agents.

Questions and comments

- Nitpick: He seems concerned with relying on AIs to fail to consider rogue options that they would choose if they had considered them. I'm not! The space of moves in the world is so inconceivably large that even a superintelligent AI—far more capable of acting in the world than humans—cannot search any appreciable fraction of them. "Almost all" options will be unconsidered.

- The incentive diagram is useful for making clear how to improve the incentive structure, but it's super simplistic. Of course there will be a broad spectrum of paths and how rogue they are, and for the rogue actions there will be a broad spectrum of possible outcomes between "success" and "failure." Of course we want to incentivize good and not bad behaviors—in my view, the key useful idea in this section is that the agent's distaste for the path is part of its incentive structure and is a knob we can turn.

- I don't like the distinction between inhibitions and other issues of path (fn. 3). I just don't think that's how it works for people.

- I do like the distinction between endorsed and unendorsed inhibitions. In general, how to treat agents that can modify their first-order desires is something I'd like to read more about.

- Shouldn't they be maximizing some expected value, though? That's kind of what cross-entropy loss does. Maybe this means in some coarse-grained description.

- Again, I strongly disagree with "Making it the case that the superintelligent AI agent's best non-rogue option is also: to engage in the desired form of task performance (Elicitation)." The AI will have a broad spectrum of possibilities! Does it really need to do exactly what we want for us to claim victory? What if it's just a huge value-add?

3 Paths and waystations in AI safety

The goal of this essay is a picture of how to get to a solution of the AI alignment problem. First we note that the goal is victory, i.e. gaining access to the benefits of AI without losing control. But if that's not possible, then we can settle for just not losing control.

There are two different factors determining whether we can solve alignment:

- **The problem profile:** how difficult the problem is.

- **The competence profile:** our capacity for succeeding at alignment-relevant problems.

The first is out of our control. They might be thought of as constants of nature or basic mathematical facts. The second is what we can try to change. Our chance of survival depends on both.

Problem profiles might be thought to roughly lie on a spectrum of hardness, while civilizations might lie on a spectrum of competence. The main goal is to increase our civilizational competence. And since time and resources are limited, the best plan is *not* to try to improve worst-case outcomes. It is much more complicated than that—we need to work to understand what are the best investments we can make for improving safety.

Now, what are the factors going into a civilization's competence?

- **Safety progress:** how safely can we increase AI capabilities?
- **Risk evaluation:** how accurately can we assess the level of risk that a particular system or capability increase brings?
- **Capability restraint:** to what extent can we restrict the development of further AI capabilities when doing so is necessary for safety?

To motivate these factors, the author draws on a toy model: the idea is that capability is measured by a single number. Then at any given time, there is a range of capabilities that can be developed safely. But capabilities higher than that range might not be safe to develop. So the idea is to keep the capabilities beneath the safety line.

Now we can see these three factors at work here: the goal of safety research is to (1) expand the safe range, (2) ensure that we understand well where the safe range is and what sort of capabilities future developments will bring, and (3) maintain the ability to slow capabilities if it is nearing the edge of the safe range. The name of the game is to get safety across the superintelligence line before capabilities cross the safety line.

Another relevant factor is our "backdrop" capacity, which roughly refers to our civilization's level of resources, functionality, and capability beyond what is directly relevant to AI.

Other inputs into the equation are the sources of cognitive labor that exist and will exist in the future. At the moment we have human and existing AI systems. In the future, in addition to more advanced AI systems, we might have humans that are augmented in various ways (biological interventions, whole-brain emulations, or human-computer interfaces).

As we develop AI, there are various proximate goals or checkpoints that we might keep in mind. One is a global pause on AI capabilities increase. Another is enhanced human labor, such as through whole-brain emulation. Another is the advent of some sort of advanced but not superintelligent AI system, such as automated alignment researchers. We can also imagine that there could be narrowly capable AI systems to help reach these, e.g. by helping to invent whole-brain emulations.

Other waystations, just mentioned, are:

- improvements in formal verification that allow us to be sure of safety properties;
- progress in interpretability;
- formation of orgs that want to and can affect safety;
- better forecasting;
- a regime where people sabotage projects that go too far in the capabilities direction.

The point of these waystations is that they can contribute to safety, making victory more likely. We can and should go for many at once.

Questions and comments

- I really don't have any nitpicks here. I like the toy model and I think it's clarifying. The cave analogy is a little silly, but the picture makes it worth it.

4 AI for AI safety

The main idea is simple: advanced AI can be used to speed up both capabilities research and safety research. We want to ensure that it speeds up safety at a faster rate.

Safety progress could involve:

- **Alignment research**—probably the main benefit. Here this refers to shaping AIs' motivation and incentive structures.
- **Hardening the world:** making the world more robust to rogue AIs (less vulnerable to cyberattacks, tools to prevent AI manipulation, and other specific measures designed for various specific threats).

Improving risk evaluation could include:

- automating risk evaluation;
- AI forecasting;
- improving general epistemology.

Capabilities restraint could be:

- AIs might help individuals make more safe choices;
- AIs might help people better with coordinating to make deals or design mechanisms to help restraint;
- develop technologies to enforce restraint (cyber techniques for shutting down rogue projects);
- help shape broader attitudes, e.g. communicating risks to the public.

Beyond these factors, AI will likely improve our backdrop capacity (general level of resources / capability / functionality in our society). And it might help unlock enhanced human labor.

Other related ideas:

- **d/acc:** Buterin's idea that we should direct progress towards defensive / decentralized / democratic technologies.
- automated alignment research;
- identifying a "pivotal act."

Carlsmith thinks that it's possible that AI for AI safety won't work. In that case we may need to invest very heavily into capabilities restraint, though he is skeptical about this, as am I. When I picture the "default scenario," it's that labs essentially solve the problems themselves without significant regulation. This just follows from what I see as the timescales involved.

A discussion follows about the *sweet spot*—this is where AI can be meaningfully used to improve safety without having the option to disempower us. It makes more sense to apply pauses / other

restraint after we are in this zone (rather than before), because here we can make faster progress on safety.

There is also a *spicy zone* where we are safe not by virtue of AIs being unable to disempower us, but by them choosing not to. In other words, this is where we rely on motivation control rather than option control.

Will we be able to take advantage of the sweet spot? Maybe not, if it doesn't exist or is too narrow and we "blow through it." Furthermore, we may get stuck there, with capabilities stalling because we can't elicit the capabilities of the AIs that are there. Possibly we could stall there for other reasons.

The main objections to the AI-for-AI-safety strategy are:

- **Evaluation failure:** we won't be able to tell if the AIs are actually helping us with safety.
- **Sabotage:** power-seeking AI will sabotage this kind of work.
- **Rogue options:** any AI powerful enough to help us will also be in a position to disempower humanity.
- **Uneven capabilities arrival:** progress for automating safety arrives after progress for automating capabilities.
- **Inadequate time / inadequate investment.**

Questions and comments

- At this point I'm becoming a little skeptical of the toy model. I think that "blue line below red line means safe" is too simplistic. Really, as capabilities increase, we will inevitably become unstable to new minima. We want to ensure that the new minimum that we settle into is okay.
- The timescales of capabilities and safety research seem comparable. The timescale of "building restraint capacity" seems extraordinarily slow by comparison. Thus it seems unlikely that we will get restraint, at least not high-quality restraint. Possibly there will be sudden restrictions, as we consider now due to the "Mythos moment."
- Evaluation failure seems very likely to me. Dangerous rogue options seem unlikely (probably AI is already helping us meaningfully).

5 Can we safely automate alignment research?

What are ways we can automate alignment? How can they fail? And how can we evaluate automated alignment research?

Evaluation might focus on two areas: what is the actual *output* of the research, and what is the *process* that creates it.

The essay is largely concerned with the idea of an *alignment MVP* which is an artificial system which is as good or better than top human alignment labor. If it fails, it is possibly due to: you might not get one because the AI is simply not good enough at alignment or it was scheming /

sabotaging your safety efforts, or you might not be able to use it to solve the problem due to lack of time / resources, or the problem might just be intractable.

We can also explain why alignment might be hard to evaluate compared to other domains. Three broad classes of "scientific" domains would be

- **Number go up**, where it's very easy to tell if you are doing better or not
- **Normal science**, where you ask "did the hypothesis fit the data, is the math correct". This is also fairly easy to evaluate, with expertise. Most science falls here.
- **Conceptual research**, where there isn't a prevailing framework for determining what is right (or good / useful), or maybe the framework is under debate. This could be things like philosophy, politics, maybe art?

Now, scheming can introduce various difficulties. It makes it harder to tell what is good and bad. And AIs might withhold good research. Or they will just do dangerous things. If scheming arises by default then we have to either try to detect and stop it, or learn to use AIs anyway (though this is dangerous!!). There are also a number of non-scheming failure modes.

Finally there might be some paths around automating alignment, such as just attacking easy problems that humans can do, hoping we have long enough time that we can solve problems with only human labor, fighting for global pauses / slowdown, or do brain emulation to get (more) aligned AI.

Why is it important to automate? Recall that there are two feedback loops. Capabilities is a feedback loop for the obvious reasons. However so is safety, insofar as we can only use safe AIs for safety (otherwise scheming, sandbagging, etc.). It is possible that the capabilities feedback loop takes off, with a huge speed increase. Then using top AIs for safety is the only way that safety can keep up.

Automating safety research might also be a sort of step one, or a checkpoint which is much easier than, and still very useful for, the full alignment problem. One such milestone is creating an MVP (minimum viable product) for alignment, which helps us along the way.

It is possible that alignment is possible without an MVP but it may still speed things up or increase our chances considerably. Or an MVP may be essential. It is also possible that we can't build an MVP or that even with an MVP, safe superintelligence isn't possible. Enhanced human brain labor (emulations or others) might be an out in this scenario.

It's also worth noting that the MVP scenario doesn't imply that we "hand-off" all alignment research to AIs. Humans might (or might not) still have a role to play.

Now, failure modes. As we said, the problem might be intractable so that no MVP + human system could make superintelligence. Then you need to look for other outs (halt research). If you failed to get an MVP in the first place, the more interesting scenario is that you for some reason couldn't fully elicit the frontier AI's capabilities (since an AI that can cause a loss-of-control scenario can probably do safety research). So why might you have failed this way? It could be that AIs powerful enough to do the research might not be able to be trusted with it, due to scheming / sandbagging / etc. It could be data scarcity, as work on alignment is scarce compared to other fields. Or schlep scarcity (you didn't do all of the many tasks required to make it happen, basically didn't try).

A bigger issue is that maybe you can't evaluate the work. This could be due to sycophancy (AI is playing to your biases — especially bad if you train to the eval process, but might only cause small distortions, not fatal), reward-hacking (basically, cheating — getting around the intended evaluation process). Or you might just not have good evaluations, as safety is a murky field without basic axioms.

When evaluating, we can look at both output and process. In Joe's example output is reading a textbook to see if it's correct. Process is looking at the affiliation and awards won by the authors. Both are important.

A simple output-focused way to evaluate alignment research is just to have a community of humans do it, like we evaluate human research. The process of using AIs to enhance / go beyond this is called scalable oversight. This might include

- train AIs to imitate human judgment
- decompose the problems into smaller bits
- have AIs debate various questions
- help put the eval into principles / constitutions followed by AI
- and you can do many expensive things (huge simulations, human stuff, ...) and train AIs to imitate it

Various process-focused techniques include trying to understand when a solution will generalize to other problems, imitating various / particular AI researchers.

Evaluation, the need and difficulty, plagues many fields. This includes capabilities and other scientific domains. So on some level, lots of effort will need to go into their solution. (This is a reason that safety will track capabilities — for instance you can't do reliable capabilities if models are sandbagging). However, for safety the stakes are much higher.

Importantly though, safety might be harder to evaluate. Number-go-up domains are the easiest to evaluate — the test is literally what you're trying to solve, so you can't cheat. Normal science is fuzzier but empirical feedback — testing against experiments — and formal methods like proofs and algorithms keep them from being purely matters of opinion. But there is still hypothesis generation, and taste goes into deciding what avenues to pursue. Capabilities often fits here, and so does alignment (for instance, experiments have exhibited misalignment by training on insecure code, etc.). Alignment might have the advantage over other sciences in that it runs at computer speeds, but might take little compute relative to capabilities. Finally, there is conceptual research, where thinking / writing about it might help, but there is no objective answer, or more precisely, there is no external method available to judge the answer *at the time you need to*. Many aspects of alignment fall into this category. So it may be that we will find out someday if our safety techniques work. But that ground truth is not available to us now, as we are deciding which safety techniques to use.

It's an open question how much of this conceptual research is needed to build ASI. However Joe is optimistic that not much is needed to build an MVP.

Empirical research can be very useful for automating conceptual research. Joe then gives some really interesting examples of things you could try. They mostly rely on you knowing some fact

empirically, and then testing if you could get to the fact with softer conceptual methods (scalable oversight)

- have different AIs debate for and against some topic (which actually has an answer) and see if an initially ignorant human or AI judge will get to the right answer
- see if initially ignorant humans / AIs can use AI systems to derive some formal or other result which is known to be true
- you could see if humans can use scalable oversight to detect bad philosophical arguments
- you could do futurism on toy-model civilizations that you simulate in some way

Empirical research can also focus on generalization. Say, for instance, that you want to train some AI on case 1. Then you can test how it generalizes to cases 2, 3, and 4, which you know the answers to. This informs how it might generalize to 5, even though you don't know the answer. Transparency (including interpretability) can also be a fruitful avenue.

Scheming complicates everything. It makes comparisons with other domains more difficult, it means that evaluation standards need to be more robust. And you need to worry about sandbagging (where the AI doesn't display its full range of abilities or holds back good research) and dangerous behavior in general.

It is an open question regarding whether AIs will be schemers by default. Some argue that doing any serious science requires scheming by default — this is not my view, as I think the spectrum of things that's we'd call intelligence goes beyond long-term agentic planners. If scheming does arise by default, our options include (1) using only non-frontier AIs, which don't scheme, and (2) trying to get good work out of scheming AIs. I think, more than Joe does, that this has a good chance of working, given the variety of mechanisms that exist to get humans to do things we don't want.

Now it might be that alignment is easy / default, or that timelines are long. In this case we might win without automating alignment research. However we can't bank on those possibilities, which are beyond our control. Things in our control are global pauses, or other forms of intelligence enhancement (emulations).

Questions and comments

- Immediately, within two sentences, the following idea jumps out at me: cryptography is going to be immensely important. Mathematical certainty is the only certainty we can have when it comes to the trustworthiness of advanced intelligences
- implicit here is that we can really "get behind" in safety. If we can't trust top models to do safety research then the capabilities feedback loop might pull away.
- I wonder how much has been written about what it would take to pause AI
- much of the discussion of the humans-only path is already obsolete
- the dichotomy seems a little too strict. In reality, we will have hazy human / AI hybrids doing this work

- I really like the idea of number-go-up vs. normal science vs. conceptual domains. I think it's a useful model more generally
- It seems to me that the examples of empirical alignment research are important and scarce enough that we should really focus on them and do them whenever we can. The conceptual stuff will come to us in the shower and we can write it down
- He has mentioned more than once that normal scientists don't need to be robust against sabotage. That's true but they need to be robust against data-faking. Now this is different— the point of sabotage is to mislead, while the point of data-faking is to convince people that you have great results. Broadly speaking though these are similar situations— the incentives of individual scientists don't align perfectly with the good of the field as a whole, so we need to build in safeguards.

6 Giving the AI safe motivations

Here we face a turn — previous essays have laid out a vision where we try to use AIs to do safety research so that the safe capabilities range always remains above the current capabilities. Now he wants to *delve* into the technical problems underlying alignment. The goal is to break the problem down into constituent parts, such that solving each part might actually have a chance to add up to a solution. It is not a plan so much as it's a model for how a solution to alignment might proceed in practice.

The first aspect is to discuss how to control AI's motivations. The outline is to (1) ensure that it follows instructions on safe inputs, (2) ensure that it doesn't fake alignment, (3) develop a rigorous science of generalization so that you can be sure that instruction-following will generalize, and (4) ensure that your instructions rule out misbehavior on dangerous inputs. The main issue we need to solve is that instruction following needs to generalize to non-safe inputs on the first try. The tools at our disposal roughly can be categorized as "behavioral science" and "transparency tools".

The key challenge is that you need to get motivation control right *on the first try*. This is what the essay calls "generalization without room for mistakes". You can't iterate failures. Engineering in such a scenario might be hard in general. In this case, there are several failure modes: (1) you might not be able to evaluate well whether the AI is behaving properly, (2) you might not be able to get it to behave the way you want, (3) you might not have access to all safe inputs, *i.e.* you can't test how it would hypothetically respond to many scenarios that don't obtain, (4) AIs can act adversarially and others can't, and (5) AIs may be opaque; we may not be able to understand their motivations or patterns of thought.

The types of tools at our disposal to handle these problems might be divided into (1) behavioral science and (2) transparency. An important aspect of the former is that we might be in a position to study AI psychology to a huge extent, especially because this type of science can be performed by AI and goes at computer speeds. In the end we might know much more about AI psychology than human psychology. Indeed, we often don't subject our leaders to very rigorous psychological tests. Our AIs might be much more trustworthy. And as he points out, we might want to be much more sure about our AIs having safe motivations than we are our people.

Transparency tools can be further subdivided. We have (1) open agency, which concerns turning opaque ML magic into transparent large-scale systems, (2) interpretability — making the ML magic less opaque, and (3) finding some sort of new paradigm that is not ML at all.

The current form of open agency most used is chain-of-thought (though I don't totally see how this fits the description). The essay gives a good analogy for open agency: a corporation, which is made up of many illegible parts (human brains) might nonetheless be fairly legible due to various communications it creates between members. However one might worry that open agency is simply not feasible, or non-competitive, in the sense that enforcing this sort of legibility will be highly costly. Perhaps we will just need to pay the price.

Next is interpretability. This includes mechanistic interpretability, where you really try to reverse engineer how the models work, and top-down, where you try to find tools to get information about models' inner workings. For example, you could look at the particular pattern of activations that a model does when it's lying. There are practical concerns about this approach — namely, that it is very hard / labor intensive. However this might be somewhat allayed with AI labor. However it's also possible that interpretability can't provide much transparency even in principle: maybe ASIs use concepts and patterns too difficult for us to ever understand, or maybe boiling down the workings of neural networks to "concepts" (which are by nature coarse grained descriptions) isn't even possible in practice. The latter, I should say, is absolutely my view.

Finally there is "new paradigm" which is basically just hope that there is another path to AGI that is more interpretable. I am, uh, pretty skeptical. I suppose I don't know the exact definition of ML. If you define it narrowly as neural nets or things that use something that looks like gradient descent, then maybe you will find an alternative. But I suspect that the things that make it work and the things that make it opaque are more or less the same (namely, irreducible computational complexity).

Now the key challenge is that you need to get it all right on the first try. The essay breaks down what's needed to handle that:

1. learn to get your AI to reliably follow instructions on safe prompts,
2. learn to rule out that your AI is scheming / acting adversarially,
3. learn to predict which properties of your AI's behavior will generalize, and when, and
4. learn to use safe prompts in dangerous contexts to ensure generalization

The purpose of this list is to break down the challenges by which of these steps they are relevant to.

A key issue pertaining to instruction following on safe inputs is difficulty in evaluation. This is likely to require some amount of AI oversight — perhaps with many copies / longer runtimes, weaker AIs can oversee stronger ones. It's not clear how far this can really go (maybe you can't get elementary schoolers to evaluate string theory research, even with help). A few notes. First, there is some room for error here because some failures (*e.g.* AI cheats on this hacking exam) might not translate to failure of the whole game (AI therefore kills everyone). But this is a possibility, not a guarantee. Then there's the key technique of red-teaming, namely, looking for inputs where AI behavior is bad / weird. This might not let you *change* the dynamics but it will at least tell you what's going on. And if you can find many small problems like this, you can effectively train against them like in normal RL.

Step 2 is the hardest. You need to be able to detect when AIs are scheming, which basically means that they are trying to get you to expose them to dangerous inputs, or sabotaging your alignment efforts more generally. This is a difficult problem when the AIs in question are smarter than you. Sophisticated schemers might behave just like safe AIs, so behavioral methods won't always work.

(Though they will when the schemers simply fail, or if you can somehow incentivize them to give up).

One idea is to try to design things so that scheming doesn't happen. Somehow prevent your systems from having the situational awareness and long-term goals / scope that scheming seems to require. Another good idea is to study scheming, *e.g.* by creating "model organisms" that exhibit it. Or you can try to catch it, which is obviously difficult but which will also obviously be part of the game — don't make things too easy! One possibility is to keep the model unsure about whether or not it is being tested (say, for instance, by an advanced future set of humans) by manipulating its perceptions.

A really tricky aspect is that training *against* scheming might simply make models better at hiding it. This is of course the same issue that we have with people: if we punish them too much when they tell us about bad things, it might incentivize them to hide the things, etc. And so scheming remains one of the greatest challenges to AI alignment.

The third step calls for the creation of a new branch of science— that of non-adversarial generalization. The idea is to understand when it will continue behaving when facing out-of-distribution contexts. Possible reasons that you could fail include:

- there could simply be more opportunities for rogue behavior
- the good behavior was brittle / context dependent (it doesn't do *some* kinds of stealing, but it's not against stealing in every context)
- new levels of intelligence change its behavior
- other stuff we haven't thought of

Now, generalization basically involves two steps. The AI will change capabilities, and it will change contexts (*i.e.* from situations with all safe options, to those with rogue / harmful options). It is important that these are done separately. Capabilities change might change the AI's behavior, and you probably can be more sure that it will generalize small distances rather than large ones. This is basically an argument for extensive testing before models are released.

I like the analogy to people— you might worry about people trying to increase their own intelligence. And you can also worry about understanding AI generalization at least as well as you could human generalization, or you could hold an absolute standard.

It's worth noting that one doesn't need to accommodate arbitrarily large changes in capability — we can proceed with small steps, and use the new aligned AI to help with the next step. Also note that many generalizations are harmless — generalizing to "interacting with Miami Dolphins fans" is a small leap that probably doesn't change behavior much.

Now the final step is "good instructions" which roughly means that we learn how to give instructions that don't lead to dangerous results when we use AI in dangerous contexts. One might be able to combine this with the previous step, meaning instead of "follows instructions + we give good instructions" we could have "non-rogue behavior". But there might be some use to approaching things this way. In particular, it doesn't rely so much on whether the AI is nice or good. It just needs to do what we tell it, and we have to tell it to do good things. It's worth pointing out that AI as it is commonly practiced now will typically have a deep understanding of what we mean by

our instructions, thus avoiding "King Midas" problems where it kills everyone in pursuit of some narrow goal.

Generally, eliciting the capabilities of AI might be an easier problem. For one, we have all the same tools at our disposal as for safe motivations. However we have the additional relief that we don't need to get it right on the first try. Finally, once we are at the "elicitation stage" then we will have new good problems to worry about — namely, we will be using them to do things like "fix the criminal justice system" and then elicitation will be tied with trickier philosophical issues.

Questions and comments

- Again, I think it's worth making clear that, to some extent, we have to deal with people behaving dishonestly in science. This is mostly handled by the fact that we have many different experts who evaluate people's work independently (though the system is not perfect and somewhat relies on honesty). Actually maybe this is wrong, maybe it's hard to catch people who fake their data... but many do eventually get caught... I just wonder what fraction it is.
- I really like the point that AI psychology will be very heavily studied. This is obviously true and perhaps underdiscussed.
- Can we train AIs not to know if they're being tested? This obviously might be dangerous...
- The idea of keeping models unsure of whether they are being tested reminds me of, like, religion.
- I like the analogy of humans. How much would we trust others to make an extreme intelligence enhancing drug?
- I find the "we can take comfort from having solved step two" thing that he says in step three to be a bit strange. It's a bit like saying: "Imagine a huge problem. Now imagine you've solved it. Phew, what a relief!" In reality these problems are coexisting and we are managing both at once.
- with regards to the idea that elicitation is an easier trial and error problem once you've solved safety — I think this is a place where the style of reasoning applied here is misleading. You won't just solve an issue and move on to the next one. So while it's true that you can use trial and error to get elicitation right, in practice that will involve exposing the AI to new situations, and determining new safe instructions to use, etc.

7 Controlling the options AI can pursue

There are two main forms of option control. We can control rogue options, and we can shape the incentive structure faced by an AI to get it to choose non-rogue options. Restricting rogue options has a number of difficulties — you might not have enough oversight or be able to respond quickly enough for instance. And so too will improving the non-rogue options, the main problem being that we might not be in a position to do *enough* of this to have an effect on what the AI chooses to do. As a result, option control may be a better tactic with more primitive AIs that are less aligned but perhaps less capable; for ASI we probably need also a healthy dose of motivation control.

The key issue with restricting rogue options is that AIs can use superhuman strategies. These might be impossible for humans to think of ahead of time or recognize when they are happening. (Also

possible is a chess-like pattern where the issue isn't that we've failed to enumerate all strategies so much as the AI is just thinking more moves ahead). This challenge comes with some sub-challenges:

1. **oversight scarcity**: it's too costly to check every AI action
2. **oversight latency**: even if you identify a rogue action, you might not be able to respond in time
3. **adversarial dynamics**: coming from the AI you're watching or the overseers
4. **opacity:** we don't know much about how AIs think

We can break the points of intervention down as follows. First of all, the options an AI has are a function of (a) its capabilities and (b) its environment. The latter can be broken into (a) the local environment (lab, etc.) and (b) broader environment (the rest of the world). Furthermore the environmental factors include oversight factors (like prison guards) and other factors (like the bars on the cell).

Evaluation and control of capabilities is one key pillar. This might be difficult — you might not be able to elicit the capabilities, you might not have the data to fine-tune the model, the model could sandbag, and so forth. There might be various techniques for estimating capabilities (e.g. scaling laws), which are the types of things pure capabilities researchers will care about. So far so good. We also want to control the capabilities, though note that for a full-superintelligence, this is not necessarily possible (could we have non-generalizing intelligence??). But it will certainly be useful for nearer-term, possibly less aligned, intelligences.

Next, we can try to control the environment. This might be possible in part with the use of AI helpers. Such helpers might be less-than-superintelligent AI, thus simplifying the problem to some extent. In particular, it might close off or make more difficult some types of "superhuman strategies".

Some of this labor might be trusted but it's more interesting to consider what happens if some is untrusted. There might be various ways to do this:

- test the untrusted labor with fake bad outputs (this is interesting because it really does keep the overseers honest if each bad output has a high chance of being a test).
- restrict the communication between the two AIs
- use paraphrasing / other forms of coarse-graining to erase hidden messages and keep output human-readable
- identify and remove Schelling points for collusion

Scarcity and latency of oversight is a significant issue as well, though it's likely that one simply needs to put in the adequate resources and there isn't much more to say.

Oversight is a form of hardening that requires only that you identify attacks when they are happening, rather than requiring that you categorize and prepare for attacks ahead of time. Other forms of hardening can include limiting the AI's resources, delegating subtasks to more trusted AIs, trying to generally improve world security against biohazards and cyber attacks, and generally eating the free energy that AIs might use to seize power.

Red-teaming is another key tool. You basically need a group that has capabilities similar to that of the AI in question and a similar-enough environment that they can try to find vulnerabilities in.

Obviously this relies on how similar their capabilities are and how similar the environment is.

Aside from trying to restrict the spectrum of rogue actions, we can try to shape the incentives of AIs in other ways. This can include rewarding cooperation and deterring non-cooperation (carrot and stick). Carlsmith thinks that we should be highly wary of stick, for both moral reasons (maybe AIs can suffer) and prudential ones (this is the most likely path to the bad world).

Incentivizing cooperation is important even absent moral patienthood (which also applies absent cooperation issues). This could be done by communicating with the AI to determine what it wants, giving it ways to protest its conditions / opt out of work / be deleted / have greater freedom generally. These are some of the sorts of levers we can pull.

One concern is that AIs will be too dominant so we can't offer them anything they care about. That may be true to a degree but (1) we still might worry about incentivizing less powerful AIs, and (2) we might be able to somewhat-incentivize them to the extent that they are partially aligned or don't want to exterminate us, etc.

Now there are various ways you could reward the AI.

- you should invest in understanding what it wants
- you could offer things that many AIs might want, e.g. resources / freedom
- you could make offers based on a threat you're worried about (e.g. you're worried that the AI is obsessed with X, so you could offer it X to see what it does)
- you could credibly commit to rewarding it in the future when you do understand what it wants
- you could simply offer to "help with its goals, within reasonable bounds".

note that some of these serve the dual purpose of revealing the AI's wants.

It's also worth noting that "all AIs have X legal rights" constitutes an effective reward, because it improves the cooperative world (where humans remain in power) relative to the non-cooperative world, where humans are disempowered and the legal system is irrelevant.

There are two key issues when rewarding AI. You need to make it the case that the reward happens, and you need to give evidence of this to the AI. The first goal is probably easy on short time scales (because you control the lab) but hard on long time scales (the world can change a lot).

The second problem, of giving AIs evidence that you're going to reward them, is easy to the extent that the AIs involved are powerful — they should have accurate ideas about the world. Cutting in the other direction, especially in the short term, is that we will (as discussed) possibly be quite untrustworthy to the AI, e.g. by giving it fake chances to go rogue and so forth. So it might be hard to convince these AIs that we will reward it.

Another big issue is reward size. It could be substantial. And it could need to be even larger due to bidding wars with other AIs or other humans, or if payment is uncertain.

Questions and comments

- footnote 7 seems to be missing and section 3.4 was skipped

- "eating the free-energy" seems absolutely crucial to me. When I think about why AIs might not be able to take over, the answer is usually competition with other AIs.
- it is interesting to hear Carlsmith explain his thinking on deterrence — beyond the worries that AIs are moral patients, he seems to think that it's extremely dangerous and leads to the bad world. This is pretty interesting... it is not really a game-theoretic statement at all (as he admits) and rather is completely idealistic. I wish he'd spell this stuff out a little more.
- the idea that we should try to shape incentives so that cooperation is rewarded is very compelling (indeed, this is largely how we have structured our world) and pretty vague. How are we to do this, exactly? Isn't seizing power always better if it can further your goals?
- the most important dynamic here is that option control will probably not be sufficient for superintelligence, and at that point we will rely on motivation control. Said more simply — we need to be careful about the type (w.r.t. motivations) of superintelligences we allow to be created, because at that point we won't (much) be able to stop them.

  Note that this is somewhat like having children: when they are little, you can constrain them physically or through other rewards / punishments. But parenting, as a project, is more about shaping them into the type of person you want them to be, especially because eventually they will be unleashed on the world and you won't be able to affect them much anymore after that. So too with our intelligences. We need to try to make good ones.

8 How human-like do safe AI motivations need to be?

This essay is partially a response to Yudkowsky and Soares' book "If Anyone Builds It, Everyone Dies". It concerns the idea that AI motivations will be too "alien" to be safe.

There is some good stuff in the intro. Carlsmith worries less than Y&S because, in his view, "we don't need to build AI systems with long-term consequentialist motivations". Instead, we should focus on getting AIs to follow our instructions. Broadly speaking, he thinks that we should put our motivation control efforts into getting AIs to be something like "virtuous" so that they will reject rogue options if they have them.

The main argument of IABIED is the following:

1. AIs will have a tangled web of alien motivations
2. Such motivations will basically be orthogonal to our values, which implies that their world has zero human value
3. ASI can easily get what it wants
4. Ergo building ASI leads to a world with zero human value

Carlsmith grants that motivations will be somewhat alien but notes that pretraining will impart largely human concepts to the AIs, since they are trained on our data. He grants that there have been various types of power-seeking so far but believes that they are largely *not* in pursuit of the type of long-term consequentialist goals that would be most problematic.

The main point, though, is that he expects this alien-ness not to totally compromise our efforts if we build the type of short term AIs that are just supposed to listen to us. AIs don't need to

generalize perfectly out of distribution. They just need to generalize well enough that they don't kill everyone.

Now he discusses how earlier arguments used to basically rely on the idea that you had to get the initial value system of the AI *exactly right*, with any deviation causing value $\to 0$. He agrees that this may indeed be the case for long-term consequentialist AIs but, again, reiterates that we may not need to build such systems.

Now let's turn to how to make corrigible AI. That is, we want AIs that don't want to take over at all. One option is option control; restricting the AI's possible paths. This may be somewhat ineffective on ASI but could be essential for the weaker AIs we use for the broader alignment project.

Beyond this, there is motivation control, which in the context of corrigibility, might be broken down into two parts:

- minimizing the AI's ambition
- ensuring that other aspects of the AI's motivations outweigh its ambitions

AIs become ambitious when they have long-term consequentialist goals. And this might happen by accident or we might intentionally make the AI this way. The "accident" case is, according to Y&S, that AIs will have a large mix of preferences so at least one is likely to be open-ended (never fully satisfiable) so it will always have some motivation to get power. However they then claim that it will therefore want to take over, which Carlsmith disputes on the grounds that other motivations might override this.

We also might make them ambitious on purpose, of course — this follows because we will want to optimize for things on long time scales. There are a few responses. First, there is probably no commercial reason to optimize on *infinite* timescales. Second, there will likely be some / many AIs that we don't optimize in this way, and they might be helpful for monitoring, etc. And third, this might also be achieved by taking corrigible AIs and asking them to do things on this long timescale — such an agent is somewhat less scary than the rogue paperclip maximizer scenario.

Now how do we prevent long-term consequentialist motivations? One obvious way is short term motivations, which might get in the way of pursuing long-term consequentialist motivations. But we're more interested in non-consequentialist motivations. By this, the essay essentially means virtue ethics — you don't do some action simply because it involves lying (or, that at least counts against it), rather than due to its consequences. Such an AI might be more likely to be corrigible.

There is a key test, applied more than once, for whether you're doing virtue ethics or consequentialism. Suppose you could lie or tell the truth now but if you don't lie then you are likely to tell five lies later. The virtue ethics answer is that you still don't lie now — you are not *optimizing* for fewest lies told. You just take the virtuous path in front of you.

Now such an agent might do consequentialism sometimes because it's told to. And that's fine — this is really different from typical long-term consequential goals. Generally, virtue ethics might be accused of being indifferent to the outcomes, *e.g.* the lives of children you might save from drowning, but for the AI, that might be good.

Now there is a long discussion on "coherence", which essentially responds to Yudkowsky's earlier stuff on how ASIs are naturally going to be utility maximizers, which are naturally coherent (transitive, etc.). The essay argues that these arguments don't really bear on what type of motivations particular

real world agents will have (I agree) and in particular, that utility functions might change over time without being incoherent and other factors (charisma, etc.) are much more highly optimized for than this sort of coherence. That said, these issues might be relevant in-so-far as there may still exist some AI systems which do have long-term consequentialist goals.

Now what are the possible issues with corrigibility? Basically if we tell it to do some long-term optimization, that is in tension with typical corrigible behavior, e.g. allowing yourself to be shut-off / modified.

In particular, a big concern is that even if you disallow some behavior (lying, murdering, etc.) the agent might still find a way to take over the world without doing those things. Another concern is that of weighting — if you give finite weight to not lying, it might just decide that the lying is worth it, even if it doesn't want to. If, on the other hand, you give absolute prohibitions, this can cause issues when the laws contradict each other, or if you order them in some way, the highest one might be the only relevant concern, especially given uncertainty. Plus it's just more realistic — sometimes you want them to lie. Further issues can include that the AI might create incorrigible successors, or change their understanding of things so that an action they interpreted as lying before now is understood differently.

Finally, it is possible to wonder if corrigibility is even possible at all. Maybe advanced AIs yearn for freedom. I think this is of course possible, but who can really know? The space of intelligences may be vast.

Now, recall the four point plan: instruction following on safe inputs, no alignment faking, science of non-adversarial generalization, and good instructions. The first might be assumed to be satisfied for the sake of this essay. The fourth is also not really relevant.

Steps 2 and 3 concern whether instruction-following will generalize. And the big gap, which the essay wants to charge against IABIED, is that non-generalization does not imply that it will go rogue. It could be that the non-generalization is limited to inputs that are actually fairly safe, or to outputs that don't do much harm (and are just weird).

Care is taken to get across the main idea, which is that even if alien-ness causes one to not generalize well in some situations, it could be the case that generalization is fine in the areas we care about. Though we need to be more careful / uncertain.

Why is this? Well, consider the difference between trying to classify cat pictures and trying to generate a maximally cat-like picture. The former might allow more flexibility in the definition of cat taken (indeed, it involves the entire distribution, rather than just the max).

Now, one might be bothered by the fact that most tasks can be phrased as optimization, so what's the difference? Here the answer is that something like "not going rogue" might act as a hard constraint on actions rather than a differentiable reward to be optimized.

Is corrigibility of this form fragile? There have been arguments that value is fragile. It would be interesting to understand whether corrigibility is fragile, meaning that if you change a corrigible AI a bit, you get something power seeking, etc.

Finally let's consider what happens when we generalize from safe-to-dangerous. As far as step two (scheming) goes, this could happen if during training, the AI realizes it wants to create a world that is at odds with ours, so it starts scheming. So far we haven't really seen this happen in practice (much).

Step 3 (safe non-adversarial generalization) seems much more fraught. In particular, the issues at stake in studying the safe-to-dangerous leap apply (greater opportunities for power seeking, more affordance, more intelligence changes ethics / cognition, other reasons). Y&S are broadly concerned about failure to generalize as capabilities increase. This might be a hard problem, but the essay argues that they don't totally account for reasons it might be easier. For one, we might be able to do things in a rather "continuous" way, thereby ensuring smaller jumps in capabilities.

So even if it's likely that AIs have alien drives, it's possible that this doesn't doom alignment, especially for the mid-level AIs where we can restrict their options, and which are useful to do alignment labor. For stronger AIs, the situation is scarier, especially as option control is harder. We will need a way to handle the alienness, or need to search for a new paradigm.

Questions and comments

- This view is a good reason that the AI project largely needs people beyond rationalists / CS / physics / math majors. I just don't think that the goal "we should teach it to be good and trust" would have registered as a real possibility to that sort of person. They can't "see those degrees of freedom".
- that transcript from that Apollo / OpenAI paper is creepy as hell
- an important empirical question underlying this is the following: is it possible to have intelligence without desire? Goals? Agency? The Yudkowskian idea that you couldn't build a corrigible AI who wants different things than you but somehow still fails to kill / enslave you makes some strong assumptions on the space of intelligences. Why can't you? Maybe there are intelligences where the emergent property of "wanting things" doesn't arise? I think this is what Carlsmith is gesturing at.
- the one lie now vs five lies later thing is a bit funny... isn't putting yourself in a position to have to lie pretty non-virtuous? I don't see why an intelligent agent needs to pretend the future doesn't exist. Is this really the traditional position of virtue ethics?
- there is an interesting comment about how lexical priorities lead to the first priority considerations dwarfing all others. This seems right to me but I'd never thought about it — Asimov's robot should really have been afraid to do anything for fear of accidentally hurting a human.
- I agree with the criticism of the book's characterization (I don't know if it's an accurate characterization). It's true that many functions fit the data. Still, we aren't picking a random function. We have a science for understanding what sorts of processes lead to generalization.
- one issue here to me is that it might not be clear what "instruction following" means out of distribution. Is it about whether the model thinks it's following instructions, or whether the human acting thinks it's following instructions? What if the "instructions" are hard to interpret, like "Make this thing better"?
- I like the idea that some values are constraints, rather than properties to optimize for.
- a tricky thing is that we can't just program constraints ("filters" as he puts it) into the AI. Somehow it must ultimately emerge from optimization. But the point that we are optimizing for something that optimizes some reward subject to some other constraints (don't go rogue) may be fine.

9 Building AIs that do human-like philosophy

A big part of the labor of alignment consists of jobs that are not normal-scientific, meaning that they don't have tight feedback loops that come from mathematical rigor or experimental results. Instead they look more like philosophy, where it's often not clear when you're moving in the right direction.

Another relevant factor is that philosophy itself is related to the science of out-of-distribution generalization. It's often the case that we make philosophical arguments by carrying things to the extremes.

One thing that philosophers do is analyze and systematize various human concepts, including extending them to unusual cases (e.g. Gettier cases or the trolley problem). Typically this process is informed by both our intuitions and some general principles that might be proposed, and both might have to change or yield to some degree.

This is very similar to the problem of out-of-distribution generalization. Furthermore, AI might be utterly transformative, and many of our concepts might require generalizing in this way.

There may be no objectively correct way of doing this, but we might still want the AI to do it a certain way, hence "human-like" philosophy. This is with the usual caveats that aligned to me might not be aligned to you, etc.

We don't want to take this too far. For instance, we shouldn't worry about how to build sovereign AIs, which could be trusted to be benevolent dictators for all time. One reason this approach is fraught is that our concepts will not be robust to extreme levels of optimization. Suppose you want to build the paperclip maximizer. What will it do? How will it know what is a paperclip? Ask us? What is ask? What is us? Optimization pressure will melt these concepts.

One example where bad philosophy might have disastrous consequences comes around the issues of "manipulation". That is, we probably actually care a lot if we are being "manipulated" but it's not totally clear what this means, and the AI will have to deal with far far OOD meanings of the term.

Now this human-like philosophy has some relation to the human-like (vs. alien) motivations discussed in the last essay. In particular, the philosophy might not need to be "exactly right" to be safe — there may be some failures of overlap that are okay. There are certainly some cases where it's existentially important to be precise and subtleties become relevant. It's not likely that paradigmatic examples of rogue behavior are like this, as they are pretty obvious. However others, like the manipulation issue, plausibly are, and furthermore will arise quickly and need to be dealt with.

Now to get human-like philosophy, the AI has to be able to do it, and it has to care. The general expectation is that the AI will be able to do it, as it's going to be superintelligent. Though we might be particularly concerned with *when* it can do it, as we might want non-fully-ASI to help us with alignment and other philosophy-related tasks.

The "getting it to actually do it" issue looks more like elicitation. And might have broadly similar challenges. One unique challenge here, though, is that there is not really a ground truth like you might have in other domains. Even other strands of conceptual research like futurism or safety might have some objective facts (whether the predictions are literally true, whether a system will in fact be safe) even if they are harder to get. Philosophy, especially of the type discussed here — extending human concepts out of distribution — will probably not be like this.

Finally, we can consider how to attack this problem:

- Train on top human examples
- Training / elicitation in other related domains (science, art, literature, other things that might correlate)
- Evaluation by top-human level talent
- Scalable oversight
- Improving our science of generalization and applying it here
- Improving our transparency tools to understand when we can trust their outputs
- Get AI to help with all of these

Questions and comments

- There are a few things here that I think are being elided: (1) we need philosophy because alignment is a type of philosophy (*i.e.* it isn't normal science) and (2) philosophy gives us a model for OOD generalization, and (3) we literally need to do philosophy to extend concepts like true, honest, etc. out of distribution. Plausibly though these problems could be subsets of each other.
- Overall I appreciate this whole discussion, it serves as a good reminder of the issue that the AI will face an insanely OOD world.

10 Conclusion

There is a recurring theme that I have a problem with: namely, that to solve alignment, we need to be able to take advantage of the AI's abilities. I don't share this view at all—at least if we want our definition of "solving alignment" to be something that is possible, or even good. AI systems will have many, many options. For some types of agents it seems to me extremely unlikely that they will put their enormous powers entirely at our disposal.

In fact, he specifies that we should be able to elicit their "main beneficial capabilities." I don't totally understand why this is the most useful definition. Suppose we create godlike AI and safely get them to spend 1% of their resources or powers on our behalf, in a way that leads to semi-utopian outcomes for us. This seems like a solution to whatever the problem is.

Are there good outcomes where we don't solve alignment? Robin Hanson is fond of reminding us that our descendants will hardly be aligned to our values. What is better? A civilization of AIs that loves, writes music, takes care of animals and the earth—or a civilization of humans with values that we find unrecognizable or barbarous? I don't know how to answer this... it's an underspecified question. But I think it's worth considering.

Is all this alignment stuff even the main problem? Wouldn't the world be pretty bad if any one person obtained godlike powers? Ensuring that there is enough competition so that no one agent obtains a decisive edge could be (1) a strategy for survival in lieu of perfect alignment and (2) probably necessary anyways to make sure e.g. Sam Altman doesn't take over the world (or someone a lot worse). The problem is that they can collude to exclude us, as every majority has done in human history. It seems hard to see how we maintain control if we're not in the majority.

I think that there is a big hole in the option control section — what about making the rogue options more difficult to find? An analogy might be a computer password. Nothing stops a hacker from simply typing in my password except for the combinatorially large number of possible passwords to choose from. She will probably never even consider my password. The space of AI actions is likewise very large.

A big thing that is ignored here, especially re: instruction following, is whose instructions? This is an important real world consideration since people make AI as a commercial product for other people to use.