Appendix 1: detailed discussion of important, actionable questions for the most important century

Questions about AI alignment

I would characterize most AI alignment research as being something like: "Pushing forward a particular line of research and/or set of questions; following one's intuitions about what's worth working on." I think this is enormously valuable work, but for purposes of this post, I'm talking about something distinct: **understanding the motivations, pros and cons of a variety of approaches to AI alignment, with the aim of gaining strategic clarity and/or changing how talent and resources are allocated.**

To work on any of the below questions, I think the first step is gaining that background knowledge. I give thoughts on how to do so (and how much of an investment it would be) in Appendix 2.

How difficult should we expect AI alignment to be?

This is a vague question. More specific versions could include “What is the expected proportion of the value of a top-tier future lost due to misaligned AI?”, “How much will AI developers need to ‘slow down’ or ‘delay’ the development of transformative AI in order to get the probability of ‘paperclipping’ below some particular level?”, etc.

Applications

In this post from the Most Important Century series, I argue that this broad sort of question is of central strategic importance. There’s a tradeoff inherent in most attempts to make the most important century go well:

- Most actions that can boost the odds of transformative AI being “in the right hands” also risk speeding the development of transformative AI, and exacerbating a “racing to develop powerful AI first” dynamic - both of which could make it harder to take intense measures toward AI alignment.
- And by the same token, most actions that can boost the odds of cooperation and caution risk *lowering* the probability that transformative AI ends up “in the right hands” (for example, because countries/labs/people who count as the “right hands” are relatively more likely to exercise caution if it is pushed for, which make them less likely to “win the race”).

There are *some* robustly helpful actions that don’t seem very sensitive to these considerations, but having a clearer best guess around various aspects of “difficulty of alignment” would open up a lot more interventions.

- If we had good arguments that alignment will be very hard and require “heroic coordination,” the EA funders and the EA community could focus on spreading these arguments and pushing for coordination/cooperation measures. I think a **huge amount of talent and money could be well-used on persuasion alone, if we had a message here that we were confident ought to be spread far and wide.**
- If we had good arguments that it won’t be, we could focus more on speeding/boosting the countries, labs and/or people that seem likely to make wise decisions about deploying transformative AI. I think a **huge amount of talent and money could be directed toward speeding AI development in particular places.**
- There are a variety of other interventions that might come from subtler points. For example, if one concluded that the alignment problem is intractable for deep-learning-based systems, one could focus on exploring other approaches to AI development and advocating against further investment in deep learning.

Example of how to attack this question

I would start by getting up to speed on AI alignment (see Appendix 2).

Then, I might:

- **Take an "anytime alignment" perspective:**
  - Write up my best shot at an "anytime alignment" strategy - a set of literal practices I could imagine taking to reduce the odds of misaligned AI in some hypothetical situation where, say, it suddenly became clear that transformative AI could be developed within a few months using today's techniques (this may not be likely, but it is imaginable).
  - My writeup would say things like "We would try to train an AI to do the following helpful task [e.g., doing a specific kind of safety research, or advising humans on strategy]; before exposing it to the full training procedure that might work for this, we would try the following set of 'sandbox experiments,' inspired by today's AI alignment agendas, and keep trying things until we saw the following sort of behavior; we would then subject the AI to broader training, while limiting its activities in the following way; etc."
  - Toss out a "probability of paperclipping-style catastrophe" conditional on this "anytime alignment" plan, get feedback from others on it, and try to elicit their reasons for disagreement and address them in my writeup.
- **Look for helpful and nonobvious (but not completely answering the question) hypotheses I could argue for**, such as "If we all we do is train on human feedback and use the best-case outputs of the following 5 safety agendas, the probability of catastrophe is very high" or "If we observe the following hypothetical experimental results, the probability of catastrophe is very low" or "If we assume that development of transformative AI is highly gradual, multipolar and CAIS-like, the probability of catastrophe is [still high, or maybe low]." I would probably write down lots of hypotheses of this basic form, start arguing for many of them in writing, discard the ones that didn't seem to be going well, and hunt around until I felt I had something I could defend.
- There are also more "socially" oriented approaches I might try, e.g. arguing with people, surveying people, or poring over public debates. I don't currently feel very optimistic about this approach on its own, but it could be a helpful supplement to the above.

Who's working on this today?

A lot of people have *opinions* on the likely difficulty of alignment, but I know of relatively few instances (certainly covering fewer than 10 people) where someone is working on an attempt to put out some research product that substantially updates its readers on the matter. And I don't expect any such products to be so definitive or comprehensive that they moot the need for more work on this, anytime in the next few years.

I expect that clarity on this question will eventually increase, as specific research agendas progress and AI capabilities advance. But (as gestured at in the next section) the situation could remain quite ambiguous for a long time by default, and having more clarity sooner could unlock interventions that would be very valuable to get started on asap.

What experimental results could give us important updates about the likely difficulty of AI alignment?

Something that bothers me a lot is that I can't easily articulate even **hypothetical** experimental results that would make me confident that an AI system is "safe," or that would substantially shift my views on the likely difficulty of alignment.

For some basic reasons this is difficult, see Why AI alignment could be hard with modern deep learning.

For now, I'd consider it progress if someone could lay out even very simplified, exaggerated, unrealistic hypothetical experimental results that many people (some pessimistic, some optimistic) agreed would be major updates in a particular direction on this point. I'd then consider it further progress if these could be refined to the point where actual experiments could be run.

Applications

If we could articulate particular experiments whose results would be informative, we could:

- Try to get actual experiments run along these lines. I'd expect this would take quite a bit of iteration and potentially a lot of money, but it would be well worth it.
- To the extent that these experiments couldn't be run yet (e.g., because AI models aren't generally capable enough yet), we could pour effort into obtaining high-quality forecasts of the results. We might start via things like Metaculus, Hypermind and Good Judgment, but I think it'd be worth **quite a bit of money and effort to augment these sorts of tools** to make them more reliable on this particular kind of question - e.g., finding ways to help forecasters gain context on AI alignment and be weighted by their understanding of it, experimenting with ways of getting better forecasts over long time horizons, etc. I am excited in the abstract about forecasting as a tool for predicting the future, but right now it's hard to apply it to any of the questions about the future I most care about; if we could make headway on having tangible, clairvoyant questions for forecasters, I think it could unlock a lot of exciting projects.
- Either way, use the results to get major updates on difficulty of alignment, and **pour money and talent into disseminating these** as in the previous section.

Example of how to attack this question

After getting up to speed on AI alignment (see Appendix 2), I would essentially:

- Write down a completely naive and not-remotely-likely-to-be-successful attempt at articulating such a hypothetical experiment, like: "Train a language model to answer questions in ways that are rated as 'helpful, honest and harmless' by human raters putting in relatively little time and effort, then see whether a set of human raters putting in more time and effort thinks these models perform well. If so, it's some evidence that

training a model to appear 'helpful, honest and harmless' is actually training it to *be* that way rather than just to trick its human raters into thinking it's that way."
- Write down a detailed explanation of what is wrong with this attempt.
- Make another attempt that addresses the previous complaint, and then write down a detailed explanation of what's wrong with this next attempt. Etc.
- As I got closer to having experiments that seemed informative, I'd put more work into reading and rereading online content about AI alignment while asking myself how much my experiments might affect the arguments being made; getting feedback from other people; etc.

Who's working on this today?

I think it is relatively common in both industry and academia for researchers to be doing something like: "Run basic experiments that are intended to gather lots of data and intuitions, without having an explicit case that they should cause major updates on the overall difficulty of alignment." I think this is very valuable work, but it is distinct from what I've described above, and what I've described above has some distinct potential applications (e.g. for funders).

Various theoretical work on AI alignment could lead to insights about this topic.

I know of fewer than 10 people who seem highly focused on this topic and/or likely to generate noticeable insights about it within the next year or so.

What relatively well-scoped research activities are particularly likely to be useful for longtermism-oriented AI alignment?

I think today's AI alignment research landscape has a lot of:

(1) Activity that fits reasonably well into existing academic and engineering traditions, but isn't necessarily aimed at the hardest and most important parts of the AI alignment problem. I think a good chunk of this activity is ultimately motivated by something more like "Improve AI systems' ability to impress human evaluators, for commercial purposes" than like "Reduce the risks of an existential catastrophe"; there is *some* degree to which the same activities are useful for both goals, but I think there are also predictable limits to the overlap.
(2) Activity that is motivated by deep concerns about existential catastrophe, but is practically incomprehensible to even very intelligent outsiders in terms of its goals, motivations, intermediate products, etc.

There is also some of:

(3) Activity that is likely to be relevant for the hardest and most important parts of the problem, while *also* being the sort of thing that researchers can get up to speed on and

contribute to relatively straightforwardly (without having to take on an unusual worldview, match other researchers' unarticulated intuitions to too great a degree, etc.)

I think **anything we can clearly identify as category (3) is immensely valuable, because it unlocks the potential to pour money and talent toward a relatively straightforward (but valuable) goal.**

Unfortunately, it is very confusing and difficult to determine what research goes in category (3). Some people think it's a very large percentage of today's AI alignment research, while others think it's a very small percentage.[1]

Working on this question could mean **arguing that a particular AI alignment agenda belongs in category (3), or coming up with a new way of thinking about AI alignment that belongs in category (3).**

Applications

I think there are a lot of people who want to work on valuable-by-longtermist-lights AI alignment research, and have the skills to contribute to a relatively well-scoped research agenda, but don't have much sense of how to distinguish category (3) from the others.

There's also a lot of demand from funders to support AI alignment research. If there were some well-scoped and highly relevant line of research, appropriate for academia, we could create fellowships, conferences, grant programs, prizes and more to help it become one of the better-funded and more prestigious areas to work in.

I also believe the major AI labs would love to have more well-scoped research they can hire people to do.

But as long as all we can say about the research we want to do/hire/support is "It should be aimed at AI safety," I think we're pretty limited. Well-scoped, identifiably crucial lines of research seem like a big bottleneck.

Example of how to attack this question

After getting up to speed in AI alignment (Appendix 2), I would start to intensely try to visualize the "best case" for most of the existing well-defined, well-resourced AI alignment tracks - in terms of what experimental results I'd expect to see, and how confident I'd feel deploying an AGI whose main hope for alignment was based on this research track (or a combination of this track and a few others). I'd list various sets of "background facts about e.g. how AI systems tend to generalize" that would imply good outcomes and bad outcomes in these hypotheticals.

[1] I tend to think that a vanishingly small percentage of today's AI alignment research goes in category (3) as I've articulated it. A noticeably greater percentage of today's AI alignment work (though still nowhere close to half) is in the category "I'm glad this exists and would be happy to fund it," while not being in category (3).

By doing this, I'd hope to get a sense of (a) which research tracks would be most valuable if they went well; (b) what the largest gaps seem to be - realistic failure modes not addressed by current lines of research, or imaginable conditions of success that a new line of research might test for - such that a new set of questions and experiments could be helpful.

I'd write up my views on (a) and (b) and seek feedback. I'd also try to backchain from these "best-cases" to intermediate observations one might expect, and try putting probabilities on those (and getting feedback on these probabilities).

Who's working on this today?

I know of very little work (certainly covering fewer than 10 people, even part-time) in the genre of "comparative alignment" - trying to argue explicitly that particular alignment agendas seem more valuable than others. (Research papers often advocate for their own approach in a section; I'm referring to work that is more explicitly focused on comparative alignment, such that it would be the main topic of a paper or report.)

For what it's worth, I *don't* think this activity is clearly valuable for leading researchers to engage in (I think it is customary, possibly for good reasons, for researchers to mostly follow their intuitions about what to work on), but I think that comparative alignment research could produce valuable products for funders and for researchers who are seeking more of a "contributor" than a "generator" role.

In terms of "coming up with new agendas," just about anyone working on AI alignment is arguably trying to produce papers, insights, and general "clarity" that could fulfill the goal here. So how you feel this is going probably comes down to what percentage of AI alignment work is in category (3) above. I will note that there are, at least, a significant number of well-informed people who seem to believe that (a) there are somewhere between zero and five people doing anything that even remotely qualifies for category (3) and that (b) we've got a big problem if nobody comes along with something a lot better than anything we've seen to date.

What's an alignment result or product that would make sense to offer a $1 billion prize for?

Is there a relatively concrete, operationalizable "alignment research result" one could describe such that:

- The result would shed a lot of light on the likely difficulty of alignment, in a way that's likely to update many people?
- And/or: the result would be valuable for improving the odds of aligned transformative AI?
- And/or: encouraging many people to energetically work toward the result would be a way of greatly improving the relevance and value of their work?

Working on this question would involve some mix of the sort of work gestured at in the previous sections, plus a lot of reasoning about psychology and social dynamics (re: how such a prize would actually affect its target audience).

Application

I think longtermist funders would be excited to fund and launch such a prize if it were well-designed. (There's nothing magic about the number $1 billion, except as an illustration that the potential impact here is high; smaller and larger prizes could be offered as well.)

I think administering such a prize would be high-stakes for many reasons other than the fiscal cost:

- It would be a play in a zero-sum competition for attention.
- It would (by design) divert people from their existing work into whatever goal is promoted by funders.
- It would be pretty unfortunate to launch such a prize and shortly afterward realize that a different prize would've been better.

I'd expect scoping the prize (to prevent spurious wins but also give maximum clarity as to the goals), promoting the prize, giving guidance to entrants, and judging entries to be a lot of work. As such, I'd be most excited to do it with a backdrop of having done the hard intellectual work to figure out what's worth rewarding.

How to attack this question

In my head, this would mostly come down to doing the sort of work described in the above sections, and picking one sort of hypothetical result or research to spend a lot of time with and turn into a potential prize. There are of course many other potential ways to generate a promising format for a prize.

Who's working on this today?

I've seen various ideas for prizes, but don't know of anyone working on the version of this I am gesturing at here, which would be largely derived from the sort of "comparative AI alignment assessment" work described in the previous sections.

Questions about AI strategy

How should we value various possible long-run outcomes relative to each other?

If transformative AI leads to lock-in, there are a number of pretty distinct-seeming outcomes:

- **Approximately as good as possible ("utopia")**, perhaps preceded by a long reflection
- **Approximately as bad as possible ("dystopia," sometimes referred to under the heading of "S-risk")**
- **"Paperclipping":** a world run by misaligned AI
- **Various "middling" worlds,** including very large but suboptimal worlds (such as we might imagine under a world government whose leaders don't end up doing much reflection and/or don't care much about general flourishing); very good but small worlds (such as we might imagine with a misaligned AI that leaves us with a small utopia, for reasons having to do with acausal trade); and robustly multipolar worlds (in which many people and/or civilizations coexist and/or compete, with no one gaining dominance).

Above, I discuss the fact that many things we might do face a potential tradeoff between reducing the odds of "paperclipping" and reducing the odds that transformative AI ends up "in the wrong hands." In order to better understand these tradeoffs, it could be crucial to have some sense of how one should value the above possible long-run outcomes relative to each other.

Most of the thinking I've seen on this topic to date has a general flavor: "I personally am all-in on some ethical system that says the odds of utopia [or, in some cases, dystopia] are approximately all that matters; others may disagree, which simply means we have different goals." I think we can do better:

- I don't think being "all-in" in the above sense is actually what the vast majority of people would endorse on reflection (even those who insist they would). I hope to expand in the future on the reasons I have this view; I am first trying to make the best case I can for an "all-in"/"bullet-biting" ethical system via my future-proof ethics series.
- Even if it makes sense to go "all-in", I think it's worth asking whether there's any way to avoid a world in which different effective altruists are at odds with each other - that is, whether some set of "compromise valuations" could be a big mutual win. I think there is also an "acausal trade" version of this same question.
- How one handles moral uncertainty could be crucial for this topic, and it seems to me that the field of moral uncertainty is extremely nascent, with lots of room to improve on existing proposals.

Application

This is a sprawling topic with a lot of potential applications. Some examples:

- A compelling "win-win" set of valuations could increase coordination and trust among the people who are both (a) interested in philosophical rigor and (b) focused on helping the most important century go as well as possible. I believe this could make a difference comparable to the "unlocking huge amounts of money and talent" ideas pointed at in previous sections.

- This question could be an important factor in many of the same tough calls listed under How difficult should we expect AI alignment to be? - insights about how we should *value* "paperclipping" vs. other outcomes could be as useful as insights about how likely we should consider paperclipping to be.
- Insights on this topic could also have more granular impacts on what sort of government, lab, etc. we should be hoping will lead the way in developing transformative AI, which could in turn unlock money and talent for making that sort of outcome more likely.

How to attack this question

I'd probably start by writing down a couple of classic "bullet-biting" answers to this question, such as "I'm a hedonic utilitarian, and I think this whole problem can be reduced to maximizing the odds of utopia." I'd then do my best to write down critiques of this position that I think would significantly update the people who take this sort of position - particularly making use of moral uncertainty and the "we should look for win-win arrangements between persons with different values" point - and try to get a sense of how compelling "bullet-biters" were finding these objections. I'd then try to "hill climb" from there: propose some approach to valuing the above outcomes that seems *better than the initial suggestion* (even if suboptimal), then point out problems with it and try to do better, etc. Along the way, I'd probably end up doing some deep dives into topics in moral uncertainty and/or decision theory.

Who's working on this today?

- Many academics think about ethics and axiology, though I'd guess that fewer than 10 are thinking about these in ways that are likely to directly and significantly inform this particular question (and I'd guess that only a handful are doing serious work on moral uncertainty in particular).
- I'm not personally aware of anyone (though I could easily be missing people) who is currently focused on driving pragmatically at a direct answer to the question I've laid out above - that is, trying to "cut through" to a best guess, with areas for further investigation highlighted, which I think may be a big value-add relative to simply working on the different pieces.

How should we value various possible medium-run outcomes relative to each other?

This is sprawling question category that could include things like:

- How much should one value "transformative AI is first developed in country A" vs. "transformative AI is first developed in country B?" That is: if transformative AI is first developed in Country A vs. Country B, how does this affect the probability of different outcomes listed in the previous section, and what ultimate implication does that have?

(This is a simplification; one can also consider things like “transformative AI is developed gradually and in a consistently multipolar dynamic” as an alternative to both, for example.)
- How much should one value “transformative AI is first developed by company A vs. company B (or first developed in a context where company A vs. B is the overall most prominent and ‘leading’ lab)?”
- How much should one value “transformative AI is developed 5 years sooner/later than it would have been otherwise?”

There are many types of subquestions that could be relevant here, including those listed in the previous section as well as things like:

- What can we infer (from history, current policy and culture, etc.) about different governments’ and societies’ general competence, likelihood of understanding and reducing catastrophic risks, and likelihood of being host to an effective “long reflection” in the event they end up dominant?
- What are the default odds that transformative AI is first developed in one country vs. another, how might these odds change in response to plausible policy changes (e.g. immigration, fab subsidies, export controls), and how do these depend on how many years it will be before transformative AI is developed?
- What are the prospects for deliberate, enforceable coordination between different governments, different labs, etc.? How likely is it that insights about AI alignment developed in one country/lab will end up applied in another?
- [luke add more?]

Application

If we were ready to make a bet on any particular intermediate outcome in this category being significantly net positive for the expected value of the long-run future, this could unlock a major push toward making that outcome more likely.

For example, funding and talent could be channeled toward building and accelerating the field of AI in a particular country (or generally), or advocating for slowing it down, or helping out companies with particular properties.

I’d guess that many of these sorts of “intermediate outcomes” are such that one could spend billions of dollars productively toward increasing the odds of achieving them, but first one would want to feel that doing so was at least a somewhat robustly good bet.[2]

[2] By this I mean not “has no downside” but rather “looks like a good bet based on the best reasoning, analysis and discussion that can reasonably be done.” I wouldn’t want to start down a path of spending billions of dollars on X while I still felt there were good extant arguments against doing so that hadn’t been well considered.

How to attack these questions

This is a particularly sprawling area for investigation - there are lots of important potential subquestions, and for each there are lots of factors one could research.

An approach we've been trying at Open Philanthropy (in work led by Tom Davidson and Joe Carlsmith) is an "integrated model," which essentially means that:

- The team created a single rough spreadsheet that tries to model a number of key factors - from the default odds of each country's having a "decisive lead" in AI development, to how these odds are affected by various interventions, to the odds of different outcomes conditional on a "decisive lead" for a particular country.
- The result is hard to follow, probably brittle (small nuances in modeling choices might make a big difference), and generally the kind of thing that I think would be rightly mocked if we were relying on it in its current form for high-stakes decision-making.
- The team is playing with different combinations of assumptions, and trying to come up with (a) conclusions that seem fairly stable across many different potential assumptions and modeling choices; (b) assumptions and modeling choices that seem particularly important and amenable to further investigation.

Who's working on this today?

There's a lot of research that could be relevant to this topic in one way or another, but in terms of direct attempts to argue something like "We should hope that transformative AI is developed under conditions like X," there's not much ongoing work I'm personally aware of outside of the attempt I mentioned above.

What does a "realistic best case transition to transformative AI" look like?

Let's presume we get something like 99th-percentile favorable "draws" for:

- How thoroughly the alignment problem will be solved by the time transformative AI is possible. (In my head, a 99th percentile outcome means it gets thoroughly solved, in that ensuring that an AI system is aligned ends up being straightforward and not very costly.)
- The extent to which those with the ability to deploy very powerful AI systems will be motivated primarily by wanting to get the best long-run outcome for the world. (In my head, a 99th percentile outcome means that people at all leading AI labs and AI-deploying companies are functionally driven by what's best for the world,[3] and that a substantial number of high-up advisors in all relevant governments are as well, although there are also a lot of other forces acting on governments.)

[3] This could be partly due to pressure from employees, investors, the public, etc. not just intrinsic motivation.

Even with these big assumptions, I still find it daunting to start picturing a step-by-step process by which we head toward a great outcome.

- If aligned AI were available as a normal commercial service, would this be conducive to great outcomes? If not, what kinds of regulations might help?
- If something like digital people became possible - digital beings that ought to have their own interests considered - who should have the right to create them, and how should voting work? (The default of "anyone can create as many digital people as they want, without limit" doesn't seem ideal, especially assuming such digital people would vote.)
- In order to facilitate a "long reflection," can and should there be a transition to some sort of global governance that can prevent different countries from racing to settle the galaxy? How could or should this transition work, and how could or should that global governance work?

I'm hoping there are some potential answers to these questions that are "cooperation-compatible." That is, I hope we can lay out a vision for how transformative AI ought to be used, such that the vision can be publicly promoted - with the result that *different players who expect each other to follow the strategy are reasonably likely to be cooperative with each other.* ("I plan to use transformative AI to forcibly establish a global government; i will make good judgment calls as I do so and afterward, but don't believe others will" is an example of a non-cooperation-compatible strategy.)

It's possible that "just ask the AI what to do" will turn out to be a perfectly fine way to handle the questions above, but I don't think we should count on that.

- For an illustration of why not, imagine that the first transformative AI is a mind upload - this would be "aligned," and a mind upload could copy itself to the point of creating a collectively "superintelligent" set of minds, but I don't think we could therefore simply "ask AI what to do" about challenges like the above.
- Furthermore, it may be wise to build relatively *narrow* transformative AI for safety reasons - AI that could help advance science and technology, and perhaps more broadly advise a human on how to achieve somewhat well-defined personal goals, but wouldn't necessarily have broad enough capabilities that we'd simply be able to defer to it on questions like the above.
- And as discussed immediately below, having a sense of what success looks like *in advance* could be important for strategy and coordination.

Application

I think that major AI labs with aspirations toward transformative AI want to "do the right thing" if they develop it, but currently have very little to say about what this would mean. They also seem to make pessimistic assumptions about what *others* would do if they developed transformative AI.

I think there's a big vacuum when it comes to well-thought-through visions of what a good outcome could look like, and such a well-thought-through vision could quickly receive wide endorsement from AI labs (and, potentially, from key people in government). If it were a cooperation-compatible vision, I imagine it could make cooperation much easier as things intensify; even if not, I think it could serve as an important North Star for making the sorts of decisions discussed in the next section.

Going from the status quo to widespread agreement on such a vision seems enormously valuable to me. I think such an outcome would be easily worth billions of dollars of longtermist capital.

How to attack this question

I'd probably start with a list of cartoonishly simplified "transition stories" for how we get from "aligned AI and well-intentioned actors" to a great outcome. Examples of "first steps" I might write down: "AI is deployed in a completely free-market way, but it all works out well because people quickly figure out how to achieve enlightenment with the help of AI-assisted technology, this becomes the most popular use of AI, and so much of the world quickly becomes enlightened and able to coordinate"; "The UN agrees in advance to regulations that its members end up implementing and defending"; "A benevolent actor forcibly establishes a world government"; etc.

I think I would then stare at my stories and ask what common factors were making most of them sound so bad to rest our hopes on, and start thinking about how these common factors might be addressed, if we could assume unrealistically large amounts of cooperation, altruism, etc. I'd gradually start to work toward at least one hypothetical story that didn't seem too absurd to tell to someone else, then start getting feedback.

Who's working on this today?

Nick Bostrom's website mentions some potentially relevant work ("doing research on the philosophy / ethics / political status of digital minds"), and a couple of his most recent papers seem relevant (here and here). Future of Life Institute is running a worldbuilding contest that could be relevant, and there are a few other people I know of exploring things like this (but I'd guess there are fewer than 10 people focused on this sort of thing overall).

How do we hope an AI lab - or government - would handle various hypothetical situations in which they are nearing the development of transformative AI, and what does that mean for what they should be doing today?

An example question in this class: "Say you run an AI lab, and you've become 95%+ confident that if you scaled up one of your systems by an amount that's affordable and would take you 3

months, you would end up with a 'virtual scientist' that could be used to automate scientific and technological advancement and quickly develop extremely advanced technologies. However, you are *not* confident that you could align such an AI - to do that confidently, you feel you would need at least another year. Furthermore, you have little sense of whether other labs are on the brink of something similar, so you're not confident that simply taking your time is the best move. What should you do? Should you build transformative AI and hope for the best? Should you approach other labs in an attempt to coordinate moving with caution, and if so what should you show them and ask them? Should you approach your government for help, and if so what exactly should you be asking of whom?"

Luke Muehlhauser and I sometimes refer to this general sort of question as the "AI deployment problem": the question of how and when to build and deploy powerful AI systems, under conditions of uncertainty about how safe they are and how close others are to deploying powerful AI of their own.

How to attack this question

(I've changed the ordering format here, because in this case I think "how to attack this question" sheds some light on "applications")

I would start with a large number of plausible hypothetical scenarios, outlining what I think an AI lab's or government's best moves would be in each, then start looking for points of commonality. I've done some of this in the past, and ended up with a simple flowchart that describes a somewhat broad class of potential deployment scenarios. One of the things I noticed from this exercise is that *information security* - keeping a trained AI's code from being stolen via cybercrime - seems both extremely difficult and extremely important in a broad range of scenarios. (This observation was part of what led to this post.) I imagine that if I kept at it, I'd generate more scenarios and notice more points of commonality.

I think the ideal way to tackle this question would involve a fair amount of background and knowledge about the technical side of AI development, such that one could have reasonable intuitions about what kind of work would be involved in developing a transformative AI model, assessing its safety, etc.

Applications

My guess is that going through exercises like the above can shed light on important, non-obvious actions that both AI labs and governments should be taking to make these sorts of potential future scenarios less daunting. Such actions could include establishing coordination mechanisms between labs and governments (I think such mechanisms should be optimized for specific purposes, and I see limited value in "coordination for the sake of coordination"); investing heavily in information security; starting to create policies and practices for determining when AI research is dangerous to publish or commercialize; and more.

As with other sections in this piece, identifying important potential actions AI labs and governments could take today could unlock interventions to encourage these actions. It could also start bringing additional clarity to questions like "Which labs, countries, etc. should we be supporting to develop transformative AI?"

Who's working on this today?

While a number of people work on topics that could have some relevance (e.g., proposing specific possibilities for AI labs to sign onto such as a windfall clause), I would say there are zero plausible-seeming proposals for "what an AI lab should generally be looking to do if it finds itself close to transformative AI," and few people (a handful at most) with a goal of generating such proposals within, say, a year.

I recognize that playing with wacky hypothetical scenarios is not a great reference class, and I believe the strategic situation will naturally become clearer if and when transformative AI draws closer, but I am really struck by how little we can say today about hopes for what "transformative AI deployment" will look like. I think small gains on this front would be extremely valuable, and I don't see them happening in the near future by default.

Questions about AI "takeoff dynamics"

To what extent should we expect a "fast" vs. "slow" takeoff?

The goal here is to improve our sense of the relative likelihood of dynamics like:

- **Dynamic 1 (extremely fast takeoff):** a self-improving AI goes, over the course of weeks or even faster, from "not very capable at all" to "so capable it can essentially take over the world on its own." In this scenario, there's no clear "fire alarm" for when transformative AI is drawing near, and even a relatively small organization with a relatively small "head start" on AI development could end up being ~all that matters.
- **Dynamic 2 (fast takeoff):** it takes something around 5-10 years to go from a world much like today's (with no universally accepted sense that transformative AI is near, steady annual economic growth of a few percent, etc.) to a world with at least 100% annual economic growth, or with some other properties that are at least as dramatically different from today's (for example, some set of misaligned AIs gaining decisive power and starting to spread through the galaxy). Under this dynamic (I am stipulating this, not claiming it follows from what I just said), a small set of companies, or a single country, could have a decisive self-reinforcing head start on AI development, and play an outsized role.
- **Dynamic 3 (slow multipolar takeoff):** we see clear, broadly accepted signs of things like "transformative AI is near or here" and "the world is dramatically changing, to a greater per-year degree than ever before" well over a decade before any sort of "point of no return." This makes it reasonably likely that there is a decade-plus period in which the

importance of the alignment problem is widely recognized *and* it's not yet too late to do important alignment work; in which the importance of the deployment problem is widely recognized *and* it's not too late to productively debate what we're hoping for; etc. Under this dynamic (I am stipulating this, not claiming it follows from what I just said), at any given point all major countries are relatively close to each other in AI capabilities, in most meaningful senses.

These are just examples, and some of the key properties can be mixed and matched.

Application

I think this question feeds importantly into a number of questions about strategy, particularly about what medium-run outcomes we should value and what sorts of things labs and governments should be prepared to do.

For example:

- Faster and less multipolar takeoff dynamics tend to imply that we should focus on very "direct" interventions aimed at helping transformative AI go well: working on the alignment problem in advance, caring a lot about the cultures and practices of AI labs and governments that might lead the way on transformative AI, etc.
- Slower and more multipolar takeoff dynamics emphasize a more "broad" approach to reducing risk of transformative AI, such as aiming to improve the general caliber of institutions.

Meaningful updates on likely takeoff dynamics could end up steering a lot of money and talent away from some interventions and towards others.

How to attack this question

One basic approach is writing down the most plausible possible detailed scenarios in which takeoff is fast or slow - including details about things like "how much more efficient AI hardware and software become as more talent and money are invested in improving them," "which AI applications can be rolled out without being bottlenecked on human-speed processes," and "the level of affordable compute that is sufficient to develop transformative AI even without much in the way of further conceptual breakthroughs." I suspect that trying to do this generates contradictions in many of the scenarios, and narrows the field somewhat. (For example, I think it is relatively hard to expect very slow takeoff if you also think that transformative AI will be developed within a few decades.)

I expect economic modeling to be a valuable tool for this work.

Who's working on this today?

Intelligence Explosion Microeconomics (Yudkowsky 2013) lays out a basic framework for tackling it, but I haven't seen much in the way of efforts to try different sets of assumptions and narrow the field of plausible scenarios. Open Philanthropy's Tom Davidson is doing work on this topic.

What are the most likely early super-significant applications of AI?

I think a common picture of how AI development will go, in this community, is something like:

- One day, AIs will be confined to fairly modest applications - exciting experimental demonstrations, some commercial uses like digital assistants and search, but nothing that is transformative for the world.
- Shortly afterward, AIs will be able to do *any* intellectual task far better than any human can, and will have the ability to completely transform and/or take over the world.

I think this is possible, but it's also **possible that AI applied in some narrow domain will be super-significant, and will massively change the world and the strategic picture before highly general AI is developed.**

For example:

- Perhaps law enforcement and militaries will rely heavily on AI-controlled robots, and this will open up different dynamics of potential coups, as humans compete to gain control of their country's military and law enforcement.
- Perhaps narrow "persuasion AIs" will be developed that can manipulate people extraordinarily effectively, at scale.
- Perhaps narrow AIs will lead to solving particular crucial scientific problems, such as being fully able to model protein-protein interaction, before they become generally capable of advancing all science.

I don't think longtermists have done much to imagine how such developments could change key strategic considerations around transformative AI, and what we could be doing today to get ahead of possibilities like those above.

Application

I think it could be beneficial for longtermists to invest in understanding, even working in, industries that might become massively more important as AI advances. By gaining experience and understanding such areas, longtermists might be in a position to help people navigate quickly-arriving, world-transforming developments brought on by AI progress, and ensure that we remain on track for good outcomes as we continue moving toward transformative AI.

I would also guess that if we could identify a few areas that seem particularly likely to see huge impact due to AI advances, this could significantly affect a number of other strategic considerations, particularly about what medium-run outcomes we should value and what sorts of things labs and governments should be prepared to do.

How to attack this question

I think this question is harder to work on independently than most of the ones listed in this document - after an initial brainstorming/canvassing period to generate potential AI applications, I'd want to talk to people who are knowledgeable both about frontier AI systems and about key industries in which it might be applied. One way to tackle this question might be to focus on creating relevant questions for forecasters (e.g., at Metaculus, Hypermind and Good Judgment), prediction markets, etc.

Who's working on this today?

I don't currently know of any ongoing work focused on this topic. There are "thought pieces" on potential future technologies, but I don't usually find them to be reasoning thoroughly about how likely various developments are, or trying to focus on the developments that could most change the strategic picture for transformative AI.

How should longtermist funders change their investment portfolios?

There are a number of ways in which a longtermist funder should arguably diverge from standard best practices in investing:

- Longermist investors should arguably have different attitudes toward risk. Exactly how one treats risk depends on the shape of the "returns curve" for longtermist philanthropy (how fast "good accomplished per dollar" declines as more dollars are spent, both overall and within a given year).
- Longtermist investors should arguably do "mission hedging": making investments such that they end up with more money in worlds where money is relatively more valuable. An example of "mission hedging" would be making investments that are likely to generate big returns if AI advances relatively quickly, since worlds where transformative AI comes soon are worlds where today's longtermist funders could have particularly outsized impacts.
- Many of the questions from earlier sections involve refining our expectations about the future, and it might be possible to "bet on" these expectations (to some degree) via investing.
- Some longtermist funders should arguably "balance out" the financial exposure of other longtermist funders, so that the pool of longtermist capital as a whole is well invested. For example, right now longtermist capital is disproportionately exposed to cryptocurrency, so some amount of hedging may be called for.

Many of the earlier sections in this piece might feed into views on investing, particularly via the first and third bullet points above.

Application

A well-argued case for making particular kinds of investments would likely be influential for major longtermist funders, collectively accounting for tens of billions of dollars in capital. On a 10-year time frame, an investment change that causes an extra percentage point of returns per year could be easily over $1 billion.

To the extent that other questions covered in this piece feed into investment decisions, this increases those questions' action-relevance and potential impact.

How to attack this question

I think this is mostly a matter of developing high-quality views on how the future is likely to play out, and what longtermist giving opportunities are likely to look like, partly via the sorts of questions discussed earlier - and combining this with knowledge of finance to produce actionable recommendations.

Who's working on this today?

Two finance professionals are in the process of starting an advisory firm that will analyze some of the topics discussed above. However, there is a huge amount of room for investigation on these topics, and I think high-quality contributions from others could make a big difference.

Appendix 2: getting up to speed on AI alignment

Here are some rough (and not always very detailed) suggestions based on the path I've been following to get up to speed. I think it's also worth checking out Richard Ngo's guide, which I recently became aware of. (I haven't tried yet to understand all of the different choices we've made and whether a consolidated guide might make sense; I may do that later, but wanted to share my own take at this time.)

A note on the "deep learning" focus here

In general, I tend to take a deep-learning-centric perspective on this topic, and don't make a lot of effort to approach the problem from a "method-agnostic" point of view (e.g., trying to say things about alignment that would apply to nearly any imaginable AI system). This is because:

- I think it's helpful to be able to think about alignment in relatively concrete and specific (if hypothetical) terms. Personally, I feel that assuming a particular "basic school of AI"

makes it much less confusing to discuss potential failure modes and solutions most of the time, though some effort to think about how one's views might generalize to other approaches is probably worthwhile.

- In practice, I think the main assumptions one takes on board in a deep-learning-centric approach are that (a) AI systems will learn to perform well on whatever tasks they're being trained on, and (b) can otherwise be considered to be more-or-less "black boxes" by default, in the sense that we can't say much by default about "how or why" they are performing well (though interpretability tools might shed light on their internals). I think these assumptions are pretty good for making oneself confront a version of the alignment problem that is both highly challenging (these assumptions mostly seem to make things harder[4] than most reasonable alternative assumptions) and pretty likely (even if there are a lot of conceptual breakthroughs between here and transformative AI, (a) and (b) seem pretty likely[5] to remain in place).
- I also think it may be reasonable to reach the conclusion that AI alignment is simply futile given these assumptions and hence move onto other approaches, but I think having a strong inside-view conviction in that conclusion, and some ability to defend it, should be achieved before making this move, and I think that requires exploring the deep-learning-centric approach pretty thoroughly.
- I think the deep-learning-centric approach is dominant at today's major AI labs, and important in academia as well. So being familiar with it is likely quite helpful for understanding what others working on AI are saying and doing. It would be even better to be familiar with a broad variety of different approaches to AI, but this seems like a good one to pick if you're picking one.

There is plenty of room for disagreement on the way I emphasize deep learning as a default framework, and there are probably many alternative versions of this appendix that could be written. I'm laying out the version that resonates with me because I think it's helpful to see an example of how one might get up to speed, even if it just inspires one to take a very different approach and emphasis.

Getting basic familiarity with today's empirical AI work

I think it's worth getting a basic sense of how today's cutting-edge deep learning systems work and how research proceeds. To this end, I recommend:

- Reading up on the conceptual basics of deep learning. I liked Michael Nielsen's conceptual explainer on neural networks and deep learning; Richard Ngo's guide has a number of other interesting-looking resources I haven't used.
- Getting some experience programming/training basic AI systems. (I haven't done this; I wish I could find the time to.) It's probably worth getting some experience with

[4] In particular, they seem conducive to a particularly large gap between the relative ease of creating transformative AI (given enough data and compute) and the relative difficulty of aligning it.
[5] Like, I'd guess least ⅓ or something, conditional on transformative AI being developed in the next few decades.

supervised learning and generative models as well as reinforcement learning (OpenAI's Spinning Up in Deep RL is potentially a good resource for the latter).
- Explainers for particularly common architectures, such as convolutional neural nets, transformers, etc.
- Reading some major papers from the last few years; one easy way to do this would be to look at releases highlighted by major labs as well as skimming "best paper" winners for major conferences (e.g.)

Reading AI alignment research

There are a lot of at-least-arguably-AI-alignment-relevant papers out there. For very wide-ranging surveys, you can check out Rohin Shah's Alignment Newsletter as well as annual discussions of AI alignment by Larks.

I don't think it's necessary to read every paper to "get up to speed." Instead, I would try to read some of the major papers - particularly those that give a lot of space to explaining their motivations and hopes - within each of the major "schools of thought" on longtermist-oriented AI alignment. A non-comprehensive list of these "schools of thought":

- **Empirical safety work being done at major AI labs**. There are probably fewer than 10 safety-focused papers authored by each of OpenAI, Anthropic, and DeepMind as of today, which can be found straightforwardly via their blogs.
- **Academic safety-oriented work.** This is a very broad area. I'd seek out papers that explicitly discuss longtermism-oriented motivations; I don't think there are many of these.
- **Work by MIRI and others focused on "agent foundations."** I found Embedded Agency most helpful; Ajeya Cotra suggests exploring Arbital.
- **Work by Alignment Research Center and Redwood Research**, which tends to be more deep-learning-oriented than the previous bullet point but more theoretically oriented and explicit in its discussion of longtermist motivations than the ones before that. I especially recommend Eliciting Latent Knowledge (and trying out the contest, even if it's over); it's straightforward to find more writing on these organizations' work via the Alignment Forum.
- A couple things that don't fit nearly in any of the above categories: mechanistic interpretability work (such as Transformer Circuits), and a couple of pieces that are explicitly "discussions of different research agendas and open problems" rather than traditional papers: Risks from Learned Optimization in Advanced Machine Learning Systems and An overview of 11 proposals for building safe advanced AI.

Aiming for deep understanding of the above

I think simply "reading" the above wouldn't suffice to be "up to speed." I'd encourage aiming for deep understanding via discussions, writing, etc.

How much of an investment is this?

I think that if someone did a significant amount of reading/tinkering/writing/understanding in all of the above categories, understood about 90%+ of the important parts of what they'd read, and had inside-view opinions on the pros, cons and overall promise of 80%+ of the above alignment agendas that tended to be considered reasonable/respectable by the authors of the papers, they'd be one of the 25 people in the world with the broadest comparative understanding of AI alignment agendas.

I don't think it would at all be easy to accomplish this (and one would need to spend a lot of time exploring things not specifically listed above, according to one's own taste). But for a very talented, deep-thinking generalist with high general technical ability but no previous experience with AI, AI alignment or even programming, I doubt it would take even 6 full-time months. That's still a huge investment - but it is probably a lot less than one would naturally expect for reaching this level of relative proficiency in such an important field.