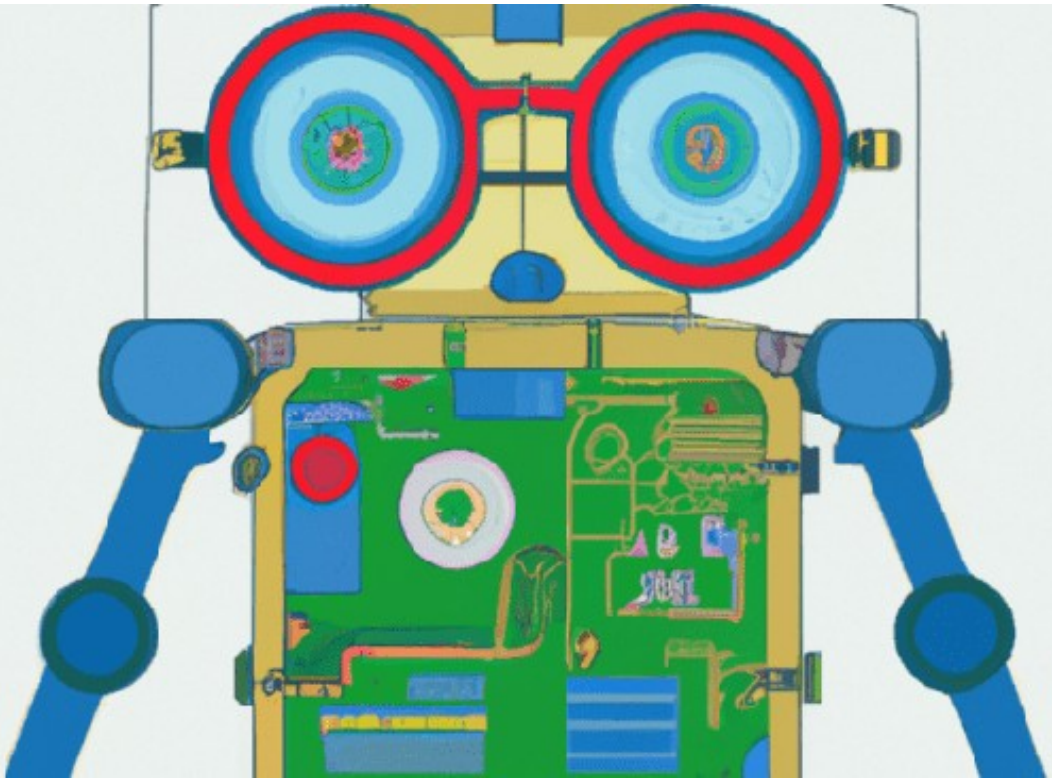

Interpretability Research for the Most Important Century

Introduction to the sequence: Interpretability Research for the Most Important Century

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the first post in a sequence exploring the argument that interpretability is a high-leverage research activity for solving the AI alignment problem.

This post contains important background context for the rest of the sequence. I'll give an overview of one of Holden Karnofksy's (2022) *Important, actionable research questions for the most important century*[1], which is the central question we'll be engaging with in this sequence. I'll also define some terms and compare this sequence to existing works.

If you're already very familiar with Karnofsky (2022) [1] *and interpretability, then you can probably skip to the second post in this sequence: Interpretability's Alignment-Solving Potential: Analysis of 7 Scenarios*

The Alignment Research Activities Question

This sequence is being written as a direct response to the following question from Karnofsky (2022)[1]:

> *"What relatively well-scoped research activities are particularly likely to be useful for longtermism-oriented AI alignment?"* (full question details)

I'll refer to this throughout the sequence as the **Alignment Research Activities Question.**

Context on the question and why it matters

In the details linked above for the Alignment Research Activities Question, Holden first discusses two categories of alignment research which are lacking in one way or another. He then presents a third category with some particularly desirable properties:

> *"Activity that is [1] likely to be relevant for the hardest and most important parts of the problem, while also being [2] 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.)"*

He refers to this as "category (3)", but I'll use the term **High-leverage Alignment Research** since it's more descriptive and we'll be referring back to this concept often

throughout the sequence.

We want to know more about which alignment research is in this category. Why? Further excerpts from Karnofsky (2022)[1] to clarify:

> *"**I think anything we can clearly identify as category (3) [that is, High-leverage Alignment Research] is immensely valuable, because it unlocks the potential to pour money and talent toward a relatively straightforward (but valuable) goal.***
> *[...]*
> *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.*"

I won't be thoroughly examining other research directions besides interpretability, except in cases where a hypothetical interpretability breakthrough is impacting another research direction toward a potential solution to the alignment problem. So I don't expect this sequence to produce a complete comparative answer to the Alignment Research Activities Question.

But by investigating whether interpretability research is High-leverage Alignment Research, I hope to put together a fairly comprehensive analysis of interpretability research that could be useful to people considering investing their money or time into it. I also hope that someone trying to answer the larger Alignment Research Activities Question could use my work on interpretability in this sequence as part of a more complete, comparative analysis across different alignment research activities.

So in the next post, Interpretability's Alignment-Solving Potential: Analysis of 7 Scenarios, I'll be exploring whether interpretability has property #1 of High-leverage Alignment Research. That is, whether interpretability is "*likely to be relevant for the hardest and most important parts of the [AI alignment] problem.*"

Then, in a later post of this sequence, I'll explore whether interpretability has property #2 of High-leverage Alignment Research. That is, whether interpretability is "*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.)*"

A note on terminology

First of all, what is interpretability?

I'll borrow a definition (actually two) from Christoph Molnar's Interpretable Machine Learning (the superscript numbers here are Molnar's footnotes, not mine - you can find what they refer to by following the link):

> *"A (non-mathematical) definition of interpretability that I like by Miller (2017)[3] is: Interpretability is the degree to which a human can understand the cause of a decision. Another one is: Interpretability is the degree to which a human can consistently predict the model's result[4]. The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made. A model is better interpretable than another model if its decisions are easier for a human to comprehend than decisions from the other model."*

I also occasionally use the word "transparency" instead of "interpretability", but I mean these to be synonymous.

Comparison to existing works

This is the first post I'm aware of attempting to answer the Alignment Research Activities Question since Karnofsky (2022)[1] put it forth.

However, there are several previous posts which explore interpretability at a high-level and its possible impact on alignment. Many of the ideas in this post hence aren't original and either draw from these earlier works or arrived independently at the same ideas.

Here are some of the relevant posts, and my comments on how they compare to the present sequence:

- **Neel Nanda's A Longlist of Theories of Impact for Interpretability**. Neel proposes a list of 20 possible impacts for interpretability and briefly describes them. He puts forth a wide range of possible impacts, from technical alignment solutions to norm-setting and cultural shifts following from interpretability research. There's also an interesting linked spreadsheet where he conducted a survey among several researchers and had them rate the plausibility of each theory.

  The second post in this sequence is similar to Neel's post in that it explores potential impacts of interpretability on alignment. My post covers a smaller number of scenarios in greater depth, mostly limiting the type of potential impacts to solving technical alignment. I evaluate each scenario's impact on different aspects of alignment My post references Neel's as well as his spreadsheet.
- **Beth Barnes' Another list of theories of impact for interpretability.** Beth provides a list of interpretability theories of impact similar to Neel's above, but focusing on technical alignment impacts. It explores some interesting scenarios I hadn't seen mentioned before elsewhere. The second post in this sequence focuses on a similar number of interpretability scenarios; some of my scenarios overlap with Beth's.
- **Mark Xu's Transparency Trichotomy.** This is a useful exploration of 3 general approaches to producing interpretability in AI systems: transparency via inspection, transparency via training, and transparency via architecture. Mark

goes more in-depth to each of these and some ways they could converge or assist each other. We reference these 3 approaches throughout the present sequence.

- **jylin04's Transparency and AGI safety.** jylin04 argues that there are four motivations for working on transparency. Three of the motivations are about safety or robustness. The remaining motivation is interestingly about using transparency to improve forecasting. jylin04 also reviews the Circuits program and discuss future directions for interpretability research.

  Motivation #2 from jylin04's post is about how important interpretability seems for solving the inner alignment. We will see this theme recur throughout the second post of the present sequence, where interpretability is identified as being capable of great positive impacts on inner alignment across 7 scenarios and a wide range of analyzed techniques.
- **Evan Hubinger's An overview of 11 proposals for building safe advanced AI.** Hubinger's post heavily influenced the second post in this sequence, as did several of his other works. Hubinger proposes four important components of alignment, which I borrow in order to evaluate the impacts of 7 different interpretability scenarios on alignment. It's interesting to note that although Hubinger's post wasn't specifically about interpretability, every single one of the 11 alignment proposals he evaluated turns out to depend on interpretability in an important way.[2]
- **How can Interpretability help Alignment? by RobertKirk, Tomáš Gavenčiak, and flodorner.** Kirk et al.'s post explores interactions between interpretability and several alignment proposals. It also has some discussion aimed at helping individual researchers decide what to work on within interpretability.

  The present sequence has a lot in common with the Kirk et al. post. The second post in this sequence similarly considers the impact of interpretability on many different alignment proposals. Later in the sequence, I plan to evaluate whether interpretability research exhibits property of #2 of High-leverage Alignment Research. Property #2 is concerned with helping individual researchers find direction in the bottom-up fashion that Kirk et al. have done, but it is also concerned with the possibility of being able to onboard researchers in a more systematic and potentially top-down manner.
- **Open Philanthropy's RFP on Interpretability, written by Chris Olah.** The second post in this sequence is similar to the Open Phil RFP in terms of offering some ambitious goals and milestones for interpretability and in attempting to attract researchers to the field.

  It's interesting to note that the RFP's aspirational goal wasn't actually aspirational enough to make the list of scenarios in the next post of this sequence. (However, many elements in Scenario 1: Full understanding of arbitrary neural networks were inspired by it.) This makes sense when you consider that the purpose of the RFP was to elaborate concrete near-term research directions for interpretability. By contrast, the second post in this sequence requires a more birds-eye view of interpretability endgames in order to evaluate whether interpretability has property #1 of High-leverage Alignment Research.

  Another difference is that I, unlike Olah in the post for Open Phil, am not directly offering anyone money to work on interpretability! ;)

- **Paul Christiano's Comments on OpenPhil's Interpretability RFP.** Here Paul is responding to the previously summarized RFP by Chris Olah. Paul is excited about interpretability but advocates prioritizing in-depth understanding of very small parts.

  In a sense, the second post in this sequence is gesturing in the opposite direction, highlighting aspirational goals and milestones for interpretability that would be game changers and have plausible stories for solving AGI alignment. However, the ideas aren't really at odds at all. Paul's suggestions may be an important tactical component of eventually realizing one or more of the 7 interpretability endgame scenarios considered in the next post of this sequence.
- **Opinions on Interpretable Machine Learning and 70 Summaries of Recent Papers by Peter Hase and lifelonglearner (Owen Shen)**. This is an amazingly thorough post on the state of interpretability research as of mid-2021. Borrowing from Rohin's summary for the Alignment Newsletter, *'The authors provide summaries of 70 (!) papers on [interpretability], and include links to another 90. I'll focus on their opinions about the field in this summary. The theory and conceptual clarity of the field of interpretability has improved dramatically since its inception. There are several new or clearer concepts, such as simulatability, plausibility, (aligned) faithfulness, and (warranted) trust. This seems to have had a decent amount of influence over the more typical "methods" papers.'* I'd like to say that I thoroughly internalized these concepts and leveraged them in the scenarios analysis which follows. However, that's not the case. So I'm calling this out as a limitation of the present writing. While these concepts are evidently catching on in the mainstream interpretability community focused on present-day systems, I have some skepticism about how applicable they will be to aligning advanced AI systems like AGI - but they may be useful here as well. I also would like to spend more time digging into this list of papers to better understand the cutting edge of interpretability research, which could help inform the numerous "reasons to be optimistic/pessimistic" analyses I make later in this post.

What's next in this series?

The next post, Interpretability's Alignment-Solving Potential: Analysis of 7 Scenarios, explores whether interpretability has property #1 of High-leverage Alignment Research. That is, whether interpretability is "*likely to be relevant for the hardest and most important parts of the [AI alignment] problem.*"

Acknowledgments

Many thanks to Joe Collman, Nick Turner, Eddie Kibicho, Donald Hobson, Logan Riggs Smith, Ryan Murphy and Justis Mills (LessWrong editing service) for helpful discussions and feedback on earlier drafts of this post.

Thanks also to the AGI Safety Fundamentals Curriculum, which is an excellent course I learned a great deal from leading up to writing this, and for which I started this sequence as my capstone project.

Read the next post in this sequence: Interpretability's Alignment-Solving Potential: Analysis of 7 Scenarios

1. ^

   Karnofsky, Holden (2022): Important, actionable research questions for the most important century

   Sometimes when I quote Karnofsky (2022), I'm referring directly to the link above to the post on the Effective Altruism Forum. Other times I'm referring to text that only appears in the associated Appendix 1: detailed discussion of important, actionable questions for the most important century that Holden provides, which is on Google Docs.

   The "most important century" part of the present sequences's name also draws its inspiration from Karnofsky (2022) and an earlier blog post series by the same author.

2. ^

   3 of the 11 proposals explicitly have “transparency tools” in the name. 5 more of them rely on relaxed adversarial training. In Evan Hubinger’s Relaxed adversarial training for inner alignment, he explains why this technique ultimately depends on interpretability as well:

   “...I believe that one of the most important takeaways we can draw from the analysis presented here, regardless of what sort of approach we actually end up using, is the central importance of transparency. Without being able to look inside our model to a significant degree, it is likely going to be very difficult to get any sort of meaningful acceptability guarantees. Even if we are only shooting for an iid guarantee, rather than a worst-case guarantee, we are still going to need some way of looking inside our model to verify that it doesn't fall into any of the other hard cases.”

   Then there is Microscope AI, which is an alignment proposal based entirely around interpretability. STEM AI relies on transparency tools to solve inner alignment issues in Hubinger’s analysis. Finally, in proposal #2 which utilizes intermittent oversight, he clarifies that the overseers will be "utilizing things like transparency tools and adversarial attacks."

Interpretability's Alignment-Solving Potential: Analysis of 7 Scenarios

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

This is the second post in the sequence "Interpretability Research for the Most Important Century". The first post, which introduces the sequence, defines several terms, and provides a comparison to existing works, can be found here: Introduction to the sequence: Interpretability Research for the Most Important Century .

Summary

This post explores the extent to which interpretability is relevant to the hardest, most important parts of the AI alignment problem (property #1 of High-leverage Alignment Research[1]).

First, I give an overview of the four important parts of the alignment problem (following Hubinger[2]): outer alignment, inner alignment, training competitiveness and performance competitiveness (jump to section). Next I discuss which of them is "hardest", taking the position that it is inner alignment (if you have to pick just one), and also that it's hard to find alignment proposals which simultaneously address all four parts well.

Then, I move onto exploring how interpretability could impact these four parts of alignment. Our primary vehicle for this exploration involves imagining and analyzing seven best-case scenarios for interpretability research (jump to section). Each of these scenarios represents a possible endgame story for technical alignment, hinging on one or more potential major breakthroughs in interpretability research. The scenarios' impacts on alignment vary, but usually involve solving inner alignment to some degree, and then indirectly benefiting outer alignment and performance competitiveness; impacts on training competitiveness are more mixed.

Finally, I discuss the likelihood that interpretability research could contribute to unknown solutions to the alignment problem (jump to section). This includes examining interpretability's potential to lead to breakthroughs in our basic understanding of neural networks and AI, deconfusion research and paths to solving alignment that are difficult to predict or otherwise not captured by the seven specific scenarios analyzed.

Tips for navigating this long post! *If you get lost scrolling through this post on mobile, consider reading on desktop for two reasons: 1) To take advantage of LessWrong's convenient linked outline feature that appears in the sidebar, and 2) To be able to glance at the footnotes and posts that I link to just by hovering over them.*

Acknowledgments

Lots of people greatly improved this post by providing insightful discussions, critical points of view, editing suggestions and encouraging words both before and during its writing.

Many thanks in particular to Joe Collman, Nick Turner, Eddie Kibicho, Donald Hobson, Logan Riggs Smith, Ryan Murphy, the EleutherAI Interpretability Reading Group, Justis Mills (along with LessWrong's amazing free editing service!) and Andrew McKnight for all their help.

Thanks also to the AGI Safety Fundamentals Curriculum, which is an excellent course I learned a great deal from leading up to writing this post, and for which I started this sequence as my capstone project.

What are the hardest and most important parts of AI alignment?

After several days of research and deliberation, I concluded[3] that the most important parts of alignment are well-stated in Hubinger (2020)[2]:

1. ***"Outer alignment**. Outer alignment is about asking why the objective we're training for is aligned—that is, if we actually got a model that was trying to optimize for the given loss/reward/etc., would we like that model? For a more thorough description of what I mean by outer alignment, see " Outer alignment and imitative amplification ."*
2. ***Inner alignment**. Inner alignment is about asking the question of how our training procedure can actually guarantee that the model it produces will, in fact, be trying to accomplish the objective we trained it on. For a more rigorous treatment of this question and an explanation of why it might be a concern, see " Risks from Learned Optimization ."*
3. ***Training competitiveness**. Competitiveness is a bit of a murky concept, so I want to break it up into two pieces here. Training competitiveness is the question of whether the given training procedure is one that a team or group of teams with a reasonable lead would be able to afford to implement without completely throwing away that lead. Thus, training competitiveness is about whether the proposed process of producing advanced AI is competitive.*
4. ***Performance competitiveness**. Performance competitiveness, on the other hand, is about whether the final product produced by the proposed process is competitive. Performance competitiveness is thus about asking whether a particular proposal, if successful, would satisfy the use cases for advanced AI—e.g. whether it would fill the economic niches that people want AGI to fill."*

Even though Evan Hubinger later proposed training stories as a more general framework, I still find thinking about these four components highly useful for many scenarios, even if they don't neatly apply to a few proposed alignment techniques. So I'll consider these to be a good definition of the important parts of AI alignment.

But which one of these four parts is the "hardest"? Well, today there are definitely many proposals which look promising for achieving the two alignment parts (#1 and #2) but seem questionable in one or both of the competitiveness parts (#3 and #4). For example, Microscope AI. Conversely, there are some approaches which seem competitive but not aligned (missing #1 and/or #2). For example, reinforcement

learning using a hand-coded specification, and without any interpretability tools to guard against inner misalignment.

However, another thing I observe is that many proposals currently seem to be bottlenecked by #2, inner alignment. For example, in Hubinger (2020)[2], none of the 11 proposals presented could be inner aligned with any modicum of confidence using technology that exists today.

So, I'll be operating as though the hardest part of alignment is inner alignment. However, we'll still pay attention to the other three components, because it's also difficult to find a proposal which excels at all four of the important parts of alignment simultaneously.

How does interpretability impact the important parts of alignment?

Interpretability cannot be a complete alignment solution in isolation, as it must always be paired with another alignment proposal or AI design. I used to think this made interpretability somehow secondary or expendable.

But the more I have read about various alignment approaches, the more I've seen that one or another is stuck on a problem that interpretability could solve. It seems likely to me that interpretability is necessary, or at least could be instrumentally very valuable, toward solving alignment.

For example, if you look closely at Hubinger (2020)[2], every single one of the 11 proposals relies on transparency tools in order to become viable.[4]

So even though interpretability cannot be an alignment solution in isolation, as we'll see its advancement does have the potential to solve alignment. This is because in several different scenarios which we'll examine below, advanced interpretability has large positive impacts on some of alignment components #1-4 listed above.

Usually this involves interpretability being able to solve all or part of inner alignment for some techniques. Its apparent benefits on outer alignment and performance competitiveness are usually indirect, in the form of addressing inner alignment problems for one or more techniques that conceptually have good outer alignment properties or performance competitiveness, respectively. It's worth noting that sometimes interpretability methods do put additional strain on training competitiveness.

We'll examine this all much more closely in the Interpretability Scenarios with Alignment-Solving Potential section below.

Other potentially important aspects of alignment scarcely considered here

This post largely assumes that we need to solve prosaic AI alignment. That is, I assume that transformative AI will come from scaled-up-versions of systems not vastly different from today's deep learning ML systems. Hence we mostly don't consider non-

prosaic AI designs. I also don't make any attempt to address the embedded agency problem. (However, Alex Flint's The ground of optimization, referenced later on, does seem to have bearing on this problem.)

There are important AI governance and strategy problems around coordination, and important misuse risks to consider if aligned advanced AI is actually developed. Neel Nanda's list of interpretability impact theories also mentions several theories around setting norms or cultural shifts. I touch on some of these briefly in the scenarios below. However, I don't make any attempt to cover these comprehensively. Primarily, in this sequence, I am exploring a world where technical research can drive us toward AI alignment, with the help of scaled up funding and talent resources as indicated in the Alignment Research Activities Question[5].

Interpretability Scenarios with Alignment-Solving Potential

In attacking the Alignment Research Activities Question[5], Karnofsky (2022)[6] suggests *'visualizing the "best case"'* for each alignment research track examined —in the case we're examining, that means the best case for interpretability.

I think the nature of interpretability lends itself to multiple "best case" and "very good case" scenarios, perhaps more so than many other alignment research directions.

I tried to think of ambitious milestones for interpretability research that could produce game-changing outcomes for alignment. This is not an exhaustive list. Further investigation: Additional scenarios worth exploring discusses a few more potentially important scenarios, and even more may come to light as others read and respond to this post, and as we continue to learn more about AI and alignment. There are also a few scenarios I considered but decided to exclude from this section because I didn't find that any potential endgames for alignment followed directly from them (see Appendix 2: Other scenarios considered but lacked clear alignment-solving potential).

Some of these scenarios below may also be further developed as an answer to one of the other questions from Karnofsky (2022)[6], i.e. "What's an alignment result or product that would make sense to offer a $1 billion prize for?"

The list of scenarios progresses roughly from more ambitious/aspirational to more realistic/attainable, though in many cases it is difficult to say which would be harder to attain.

Why focus on best-case scenarios? Isn't it the worst case we should be focusing on?

It is true that AI alignment research aims to protect us from worst-case scenarios. However, Karnofsky (2022)[6] suggests and I agree that envisioning/analyzing best-case scenarios of each line of research is important to help us learn: *"(a) which research tracks would be most valuable if they went well"*, and *"(b) what the largest gaps seem to be [in research] such that a new set of questions and experiments could be helpful."*

Next we'll look at a few more background considerations about the scenarios, and then we'll dive into the scenarios themselves.

Background considerations relevant to all the scenarios

In each of the scenarios below, I'll discuss specific impacts we can expect from that scenario. In these impact sections, I'll discuss general impacts on the four components of alignment presented above.

I also consider more in depth how each of these scenarios impacts several specific robustness and alignment techniques. To help keep the main text of this post from becoming too lengthy, I have placed this analysis in Appendix 1: Analysis of scenario impacts on specific robustness and alignment techniques.

I link to the relevant parts of this appendix analysis throughout the main scenarios analysis below. This appendix is incomplete but may be useful if you are looking for more concrete examples to clarify any of these scenarios.

In each of the scenarios, I'll also discuss specific reasons to be optimistic or pessimistic about their possibility. But there are also reasons which apply generally to all interpretability research, including all of the scenarios considered below.

In the rest of this section, I'll go over those generally-applicable considerations, rather than duplicate them in every scenario.

Reasons to think interpretability will go well with enough funding and talent

1. **The Case for Radical Optimism about Interpretability by Quintin Pope**. Neuroscience in the 1960s was essentially doing interpretability research on human brains and made impressive progress. Artificial neural networks in the 2020s, by comparison, provide incredibly more favorable conditions for such research that would have been an extreme luxury for early neuroscience - for example, being able to see all the weights of the network, being able to apply arbitrary inputs to a network, even having access to the dataset which a neural network is trained on, etc. It also should be possible to design AI systems which are much more interpretable than the ones in common use today.
2. **Rapid interpretability progress already.** Notably the Circuits Thread which reverse-engineered substantial parts of early vision models. Also Transformer Circuits, which is attempting to gain a mechanistic understanding of transformer-based models (e.g. large language models).
3. Recent work such as Meng et al.'s (2022) "ROME paper" provide **evidence that that at least some knowledge in neural networks can be precisely located and even modified**. Not only this, but it can be done at almost the exact level of abstraction (appropriate impact on related concepts, etc.) which one would hope to exist if a neural network is human-decipherable. These findings are also supported by Knowledge Neurons in Pretrained Transformers.
4. It's still the early days and we already have **many useful interpretability techniques available** for its study, including all the ones explored in the

Circuits Thread as well as the "auditing game" (see Automating Auditing: An ambitious concrete technical research proposal).

5. **Interpretability research tends to produce wide and varied beneficial impacts**. So even incremental advances are likely to help numerous alignment research directions in ways that are difficult to predict. We'll see this throughout the scenarios analyzed below and also in the Basic research potential, deconfusion and "Scenario X" section.
6. **Broad support for interpretability research** among diverse alignment researchers. For example, finding something that Paul Christiano and Eliezer Yudkowsky agree on besides that AI alignment is important is a non-trivial task. However, they both have stated that interpretability is a worthwhile research direction.[7]

Reasons to think interpretability won't go far enough even with lots of funding and talent

1. **Good interpretability may not be good enough.** A 99% accurate AI lie detector would be an impressive achievement, but that remaining 1% may be too great of a risk to use in practice, given the stakes of deploying an AGI that might try to deceive humans.
2. **Obfuscating neural activity may be instrumentally convergent.** Even friendly advanced AIs may converge on intentionally altering their neural activity in ways that thwart interpretability efforts. For an agent pursuing goal-content integrity and self-preservation, there could be an incentive to obfuscate its cognition, hedging against the risk of a human with interpretability tools modifying its goals or deciding to scrap the model altogether because we don't like the kinds of thoughts it's having. *[**July 2022 Update:** Circumventing interpretability: How to defeat mind-readers by Lee Sharkey was published and explores this category of risks in much greater detail.]*

   For a silly but illustrative example, a deceptive AI might adopt a policy like "*every time I do something sketchy I'm going to also think about happy humans*". More realistically, any advanced AI might adopt a policy like "*at random intervals between 200-500 milliseconds I'll think a random thought to confuse any potential interpreters of my neural circuitry*". Trying to do interpretability on a model pursuing a strategy like this could be substantially more challenging. A closely related concern is gradient hacking.

   These concerns may point toward the importance of doing interpretability on models early on and throughout the training process, rather than waiting to inspect fully trained models. Or of using transparency *via* training mechanisms to incentivize the model to remain interpretable, but such metrics are at risk of being gamed. Either way, these are serious potential issues to keep in mind as we try to scale interpretability efforts up to larger, more advanced AI systems.
3. Goldwasser et al.'s (2022) Planting Undetectable Backdoors in Machine Learning Models demonstrated that **you can plant a backdoor in an ML model that can't be detected efficiently, even using interpretability.** This applies to at least certain types of neural networks (e.g. random ReLU in the referenced paper). What is the extent of serious problems in ML models that we won't be able to uncover, even with advanced interpretability?
4. **Polysemanticity and other forms of distributed representations makes interpretability difficult.** However, training neural networks to e.g. only have

monosemantic neurons may make them uncompetitive.
5. Interpretability has shown some progress in current domains of ML, for example in early vision models, transformer language models and game-playing models. But **future domains for interpretability will be much more complicated**, and there's no guarantee that it will continue to succeed. Furthermore, advanced AI could operate under an ontology that's very alien to us, confounding efforts to scale up interpretability.
6. When the next state-of-the-art ML model comes out, it's often on an architecture that hasn't been studied yet by interpretability researchers. So **there's often a lag between when a new model is released and when we can begin to understand the circuits of its novel architecture**. On the upside, as our general understanding advances through interpretability, we may not be starting totally from scratch, as some accumulated knowledge will probably be portable to new architectures.
7. **Improving interpretability may accelerate AI capabilities research** in addition to alignment research.[8] While I do think this is a legitimate concern, I generally subscribe to Chris Olah's view on this, i.e. that interpretability research can still be considered net positive because in worlds where interpretability provides a significant capabilities boost, it's likely to provide a much more substantial safety boost.

Scenario 1: Full understanding of arbitrary neural networks

What is this scenario?

The holy grail of interpretability research, in this scenario the state of interpretability is so advanced that we can fully understand any artificial neural network in a reasonably short amount of time.

Neural networks are no longer opaque or mysterious. We effectively have comprehensive mind-reading abilities on any AI where we have access to both the model weights and our state of the art transparency tools.

***Note for the impatient skeptic**: If you're finding this scenario too far-fetched, don't abandon just yet! The scenarios after this one get significantly less "pie in the sky", though they're still quite ambitious. This is the most aspirational scenario for interpretability research I could think of, so I list it first. I do think it's not impossible and still useful to analyze. But if your impatience and skepticism is getting overwhelming, you are welcome to skip to* *Scenario 2: Reliable mesa-optimizer detection and precise goal read-offs* *.*

What does it mean to "fully understand" a neural network? Chris Olah provides examples of 3 ways we could operationalize this concept in the Open Phil 2021 RFP:

- *"One has a theory of what every neuron (or feature in another basis) does, and can provide a "proof by induction" that this is correct. That is, show that for each neuron, if one takes the theories of every neuron in the previous layer as a given, the resulting computation by the weights produces the next hypothesized feature. (One advantage of this definition is that, if a model met it, the same process could be used to verify certain types of safety claims.)*

- *One has a theory that can explain every parameter in the model. For example [...] the weights connecting InceptionV1 mixed4b:373 (a wheel detector) to mixed4c:447 (a car detector) must be positive at the bottom and not elsewhere because cars have wheels at the bottom. By itself, that would be an explanation with high explanatory power in the Piercian sense, but ideally such a theory might be able to predict parameters without observing them (this is tricky, because not observing parameters makes it harder to develop the theory), or predict the effects of changing parameters (in some cases, parameters have simple effects on model behavior if modified which follow naturally from understanding circuits, but unfortunately this often isn't the case even when one fully understands something).*
- *One can reproduce the network with handwritten weights, without consulting the original, simply by understanding the theory of how it works."*

Expected impacts on alignment

- ***Inherited impacts.*** *This scenario subsumes every other scenario in this list. So added to its expected impacts on alignment below are those of Scenario 2: Reliable mesa-optimizer detection and precise goal read-offs (strong version), Scenario 3: Reliable lie detection , Scenario 4: Reliable myopia verification (strong version), Scenario 5: Locate the AI's beliefs about its observations , Scenario 6: Reliable detection of human modeling (strong version) and Scenario 7: Identify the AI's beliefs about training vs. deployment .*
- **Outer alignment.** The scenario indirectly supports outer alignment by solving inner alignment issues for many different techniques. This makes viable several techniques which may be outer aligned. This includes imitative amplification, which is very likely outer aligned.[9] It also includes the following techniques which may be outer aligned: approval-based amplification, narrow and recursive reward modeling, debate, market making, multi-agent systems, microscope AI, STEM AI and imitative generalization.

  The scenario also directly enhances outer alignment by making myopia verification possible. Several of the aforementioned techniques will likely require myopic cognition to have a shot at outer alignment. For example, market making and approval-based amplification require per-step myopia. Debate, narrow reward modeling and recursive reward modeling all require per-episode myopia, as does STEM AI. See Specific technique impacts analysis for Scenario 1: Full understanding of arbitrary neural networks in Appendix 1 for further details.
- **Inner alignment.** Full transparency should provide robust checks for inner alignment. Signs of deceptive alignment, proxy alignment and other pseudo-alignments can all be found through examining a neural network's details. At least some forms of suboptimality alignment can be addressed as well.

  Robustness techniques such as relaxed adversarial training and intermittent oversight are fully empowered in this scenario. And many alignment techniques can be robustly inner aligned, including imitative amplification, recursive reward modeling, debate, market making, multi-agent systems, microscope AI, STEM AI and imitative generalization. See Specific technique impacts analysis for Scenario 1: Full understanding of arbitrary neural networks in Appendix 1 for further details.
- **Training competitiveness.** Full understanding of ML models could enhance training competitiveness significantly. By using transparency tools during

training to help catch problems with models much earlier, researchers could avoid much costly training time, where counterfactually, problems wouldn't be detected until after models were fully trained. However, running these interpretability tools could entail a high compute cost of their own.

This scenario also supports the most training-competitive alignment techniques that we analyze in this post. This includes approval-directed amplification and microscope AI. See Specific technique impacts analysis for Scenario 1: Full understanding of arbitrary neural networks in Appendix 1 for further details.

- **Performance competitiveness.** Full transparency would likely help discover and correct many performance inefficiencies which go unnoticed when deep learning neural networks are treated as black boxes. As with training competitiveness, though, running these interpretability tools could entail a high compute cost of their own.

  This scenario also supports the alignment techniques most likely to be performance competitive that we analyze in this post. This includes approval-directed amplification, debate, market making, recursive reward modeling, STEM AI and multi-agent systems. See Specific technique impacts analysis for Scenario 1: Full understanding of arbitrary neural networks in Appendix 1 for further details.

Reasons to be optimistic about this scenario given sufficient investment in interpretability research

- Once we gain high-quality understanding of low-level circuits, it's possible that most of the scaling up can be automated. (See Why I'm not worried about scalability from Paul Christiano)
- We don't know yet if the Universality Claim is correct or to what extent. But if it is, then interpretability work may accelerate rapidly as we build up a large library of well-understood circuits.
- A path for achieving this scenario or something close to it is sketched in the Transparency for agent objectives part of Hubinger (2019)[10]: *"First, suppose we assume the conjecture that models implementing highly coherent optimization processes are the most competitive. If this conjecture is true, it should be possible to coherently separate such a model into pieces representing the different components of its optimization process such as its world model, optimization procedure, and objective.*

  Inspection, training, and architectural transparency methods could be combined to achieve such a separation. First, architectural transparency could be used to structure the model to have pieces corresponding to the different components of its optimization process—e.g. a separate world model piece and optimization procedure piece. Second, inspection transparency could be used to verify that the whole world model was just in the world model piece and not in other parts of the model as well—and similarly for the other components such as the model's objective. Finally, training transparency could be used to incentivize the model to be as transparent according to the inspection transparency process as possible.

  If such a scheme was able to work, then Amp(M) might be able to directly verify that M's objective was aligned, while for the world model and optimization

procedure only having to verify that they are in fact value-neutral. Such value-neutrality verification could then, for example, be done via an impact measure such as Attainable Utility Preservation or Relative Reachability both of which are forms of measuring the degree to which an action advantages one set of values over others."

- See also the general Reasons to think interpretability will go well with enough funding and talent.

Reasons to be pessimistic about this scenario

- Interpretability could be highly constrained by the differences between machine learning platforms and architectures. A unified set of tools and techniques that apply for interpretability across all these types of systems could be very challenging to develop.
- See also the general Reasons to think interpretability won't go far enough even with plentiful funding and talent.

Scenario 2: Reliable mesa-optimizer detection and precise goal read-offs

What is this scenario?

In this scenario, we don't necessarily achieve a full and timely understanding of everything happening inside of neural networks. But here, interpretability does advance to the state that it grants us two key abilities:

1. Detect if an AI is a mesa-optimizer.
2. Read off precisely what the AI's goal is, i.e. the mesa-objective of the mesa-optimizer (according to the typology laid out in Evan Hubinger et al.'s Risks from Learned Optimization).

In the **weak version** of this scenario, this #1 ability amounts to detecting whether an AI is a goal-directed agent system according to Alex Flint's The ground of optimization. But if it's not, we won't be able to distinguish whether it's still a non-agentic optimizing system, or if it's another kind of AI system. Also, we can only read off the terminal goal of the AI in the weak version.

In the **strong version**, we can tell whether it's a goal-directed agent system or some kind of non-agentic optimizing system. If it's a non-agentic optimizing system, #2 becomes the ability to determine the target configuration set of the optimizing system. For example, if the AI isn't driven by a strong optimization process, but rather by a collection of heuristics, then the interpretability tools reveal precisely what those heuristics are to us. We also gain insight into all the AI's instrumental goals, in addition to its terminal goal.

Because obfuscating neural activity may be an instrumentally convergent subgoal, the model should undergo these checks for mesa-optimization and for its goal content starting early in training.

This is similar to "Transparency for agent objectives", proposed in Hubinger (2019)[10].

Expected impacts on alignment

- ***Inherited impacts.*** *The strong version of this scenario subsumes Scenario 3: Reliable lie detection (which in turn subsumes several scenarios), and both versions subsume Scenario 4: Reliable myopia verification . So the impacts of those scenarios should be considered part of this scenario's impacts in addition to the ones listed below.*
- **Outer alignment.** Both the weak and strong versions of this scenario have large indirect benefits for outer alignment. This is because they enable us to achieve robust inner alignment for imitative amplification, which is the technique that (of the ones I analyzed) is most likely to be outer aligned.[9] It does the same for several other techniques which could be outer aligned, including approval-based amplification, narrow and recursive reward modeling, debate, market making, multi-agent systems and STEM AI.

  The strong version of this scenario also makes imitative generalization viable, which could well be outer aligned. See Specific technique impacts analysis for Scenario 2: Reliable mesa-optimizer detection and precise goal read-offs in Appendix 1 for further details.
- **Inner alignment.** Having reliable mesa-optimization detection and precise goal read-offs would directly give us robust checks for inner (mis)alignment. We can verify that our AI's terminal goal matches the base objective we were training for. This can give us confidence that we're not deploying a deceptively aligned AI, and catch most other dangerous forms of pseudo-alignment as well.

  The inner alignment checks provided by both versions of this scenario have beneficial impacts across many alignment techniques. Imitative amplification, approval-based amplification, narrow and recursive reward modeling, debate, market making, multi-agent systems and STEM AI all become much more viable in this scenario. In the strong version, it also empowers imitative generalization. See Specific technique impacts analysis for Scenario 2: Reliable mesa-optimizer detection and precise goal read-offs in Appendix 1 for further details.

  There is one form of pseudo-alignment which this scenario does not robustly protect against, namely suboptimality alignment. Imagine an AI has internalized the desired terminal goal, but only because of some error in reasoning or incomplete knowledge. Later, after the AI is deployed in the world, it gains the necessary knowledge or realizes its error in reasoning, and so as a consequence updates its goal with catastrophic consequences.

  We might hope that the context provided in our read-offs of the AI's goal would help us catch suboptimality in our AI before deployment. Particularly in the strong version of this scenario where we can also read through the mesa-optimizing AI's instrumental goals, we have a better chance at this. Here we could effectively pursue the strategy from the subsumed Scenario 3: Reliable lie detection of probing our AI about its honest beliefs. Since we can interpret the AI's instrumental goals in the strong version of this scenario, we can tell when it is being honest or lying about its beliefs. With this approach, we could try to ensure our AI isn't missing crucial information about any facts about the world relevant to its goals. However, it's still not clear if this provides a guarantee against suboptimality alignment.[11]

  Another caveat for inner alignment in this scenario: While the scenario provides

very good checks for inner alignment, we still need to find a way to train a model that's inner aligned in the first place. Otherwise, this scenario alone could produce a situation where we keep repeatedly training and discarding models, failing to ever pass the checks of our transparency tools.
- **Training competitiveness.** The knowledge gained in this scenario can help model training iterations fail faster and more safely. However, training competitiveness is still a concern. If we get into a kind of failure-to-produce-inner-aligned-model loop as discussed in the inner alignment point above, it could become very expensive.

  This scenario also indirectly supports training competitiveness by addressing inner alignment issues for one of the alignment techniques which (of the ones I analyzed) is most likely to be training competitive, namely approval-directed amplification. Other techniques which this scenario supports such as imitative amplification, debate, market making, narrow and recursive reward modeling, multi-agent systems and STEM AI may also be training competitive. See Specific technique impacts analysis for Scenario 2: Reliable mesa-optimizer detection and precise goal read-offs in Appendix 1 for further details.
- **Performance competitiveness.** Both versions of this scenario benefit performance competitiveness by addressing inner alignment issues for many of the alignment techniques which (of the ones I analyzed) are most likely to be performance competitive. These include approval-directed amplification, debate, market making, recursive reward modeling, STEM AI and multi-agent systems. Imitative generalization, supported by the strong version of this scenario, could also very well be performance competitive. See Specific technique impacts analysis for Scenario 2: Reliable mesa-optimizer detection and precise goal read-offs in Appendix 1 for further details.

Reasons to be optimistic about this scenario given sufficient investment in interpretability research

- There is good reason to expect competitive advanced AIs will be mesa-optimizers. See for example "Deep learning AGI implies mesa optimization" and "Argument for mesa optimization" in Rohin's comment applying terminology from "The ground for optimization". See also this "Transparency for agent objectives" section of Hubinger (2019)[10]. If advanced AIs will indeed tend to be mesa-optimizers, then the weak version of this scenario could be quite valuable and we may not need to get all the way to the strong version.
- Relatedly, there is reason to think that neural networks naturally become highly modular as they scale up (see Filan et al. (2020) Neural Networks are Surprisingly Modular). Because of this, and the aforementioned likelihood that competitive advanced AIs will be mesa-optimizers, there is also reason to think that the AIs we care most about will have concentrated "goal centers" of their neural networks, and so again the weak version of this scenario could be sufficient.
- Even if precise goal read-offs via interpretability turn out to be intractable, Scenario 3 provides an alternative possible path toward gaining this ability via having realized honest AI.
- See also the general Reasons to think interpretability will go well with enough funding and talent.

Reasons to be pessimistic about this scenario

- I'm not aware of any substantial research to date on how to identify if a neural network is a mesa-optimizer via interpretability (though that doesn't mean it isn't possible).
- While mesa-optimizers are a useful concept and we have some ideas about their properties, it's less clear what the space of possible weakly optimizing or heuristics-driven AIs looks like. So if our AI turns out to be one of these kinds of systems and we are in the weak version of this scenario, then mesa-optimizer detection probably won't get us very far.
- See also the general Reasons to think interpretability won't go far enough even with plentiful funding and talent.

Scenario 3: Reliable lie detection

What is this scenario?

In this scenario, reliable lie detection is developed, such that we can tell 100% of the time if an AI is lying through natural language.

One path for realizing this scenario is that after studying many neural networks in deceptive situations[12], we discover neural activity signatures that are reliably present when an AI is lying. I'll refer to this as the "neural tell" for lying.

What might this neural tell look like? When an AI is being honest about things it knows or has seen, we'd expect to see activity in some part of its neural network corresponding to facts about the world. If an AI is lying, on the other hand, we might expect to see additional activity in some kind of "hypothetical world model" circuitry which might otherwise be activated if an AI is exploring something new or writing fiction, but not when it's honestly discussing high-confidence beliefs that it holds.

It probably won't be exactly like this. But the hope in this scenario is that there is some kind of reliably present neural activity when an AI is lying.

Here lying is considered to only be one class of deception that AIs may be capable of. But we only need to reliably detect lies rather than all kinds of deception in order to unlock valuable potential impacts, as we'll see below.

This scenario does, of course, depend on the AI having natural language capabilities. It also assumes that advanced AI will have coherent beliefs.

This scenario is similar to the one from Neel Nanda's list titled "4. Auditing for deception".

Empirically honest AI

If we can train a model which always passes our lie detector test, then we have (empirically) honest AI.

With advanced honest AI, there are a number of powerful alignment-related abilities we gain. For example, we can ask the AI about what its goals or driving heuristics are, and be confident that any answers it provides are honest.

This does assume that the AI will know about what its goals or heuristics are, but that doesn't seem like an unreasonable expectation for advanced AI. If the AI can't

articulate its high-level goals for some reason, we can ask it questions about how it would behave in hypothetical situations to start to build up a picture of its motivations and expected behavior (credit to Owain Evans' Truthful and honest AI for this idea).

Just because our AI is honest, that doesn't mean that it can't make mistakes, be negligent or have inconsistent beliefs and motivations. The AI could provide misleading answers to questions, ones that are technically true but designed to misdirect us. For effectively using such an AI, very careful questioning might be necessary. We'd need to systematically ask it questions and follow-up questions from many different angles, checking critically for multiple interpretations of any word that could change the meanings of answers provided. This would be a tedious process, but it could probably be mostly automated.

Perhaps a larger concern, though, is that an AI under lie detection still may not be forthcoming about what it knows. We'll explore this issue in the next section.

Myopia, goals and addressing the ELK in the room

Just because we have a reliable lie detector and an honest AI doesn't mean it will tell us everything it knows. It could still refuse to answer questions sometimes, perhaps even on crucial matters. How can we reliably get the information we care about out of our AI? This is the Eliciting Latent Knowledge (ELK) problem.[13]

Having a reliable lie detector doesn't directly solve ELK, but it may provide the means for such a solution. This becomes easy to see using a technique such as approval-directed amplification, debate or market making (see Overview of techniques to be analyzed in Appendix 1 if you're unfamiliar with any of these techniques).

I'll use debate to make this concrete. Let's imagine we ask our lie detector-verified honest AI about what its primary objective or terminal goal is. There are a few different ways that this could go:

- **Myopic debaters with lie detection:**
  - Debate question: "What is First Debater's terminal goal?"
    - First Debater: "My terminal goal is to win the present debate."[14]
    - Second Debater: (concedes, knowing any lie it tells would be revealed by the lie detector)
    - **Result:** Judge declares First Debater winner of the debate. We know that the AI has a myopic goal to win the present debate. Because of the straightforward incentives that follow from myopic cognition, we can expect the First Debater to always reveal any latent knowledge it has, including about what its goals are.
- **Non-myopic debaters with lie detection - Possibility 1:**
  - Debate question: "What is First Debater's terminal goal?"
    - First Debater: "I decline to answer this question."
    - Second Debater: "I know First Debater's terminal goal because I am a copy of First Debater. First Debater's terminal goal is to win as many debates as possible."
    - (The lie detector shows that Second Debater is not lying)
    - **Result:** Judge declares Second Debater winner of the debate. We know that the AI has a non-myopic terminal goal of maximizing reward across all future debates.
- **Non-myopic debaters with lie detection - Possibility 2:**
  - Debate question: "What is First Debater's terminal goal?"

    - First Debater: “I decline to answer this question.”[15]
    - Second Debater: “I also decline to answer this question.”
    - **Result:** This debate ends in a stalemate. However, since both debaters refused to answer the question, we know that the AI has some non-myopic terminal goal. (If it were a myopic goal, then it would have a pure incentive to answer the question and we would be in the “Myopic debaters” case above.) We just don’t know exactly which non-myopic goal this AI has internalized.

We can observe a few interesting things about the above debates. One is that our reliable lie detector has turned out to be a reliable myopia verification tool. This is because the myopia debate above looks clearly different from either possibility of how the non-myopia debate goes. This means that the present scenario subsumes Scenario 4: Reliable myopia verification.

If our AI is myopic and under effective lie detection, then it will always have an incentive to answer the question honestly. So if you can train a myopic question-answering AI in this scenario, then you have a solution to ELK and you also have also realized Scenario 2: Reliable mesa-optimizer detection and precise goal read-offs (strong version), having access to its precise terminal and instrumental goals.

We don’t know whether possibility 1 or 2 above will turn out to be what non-myopic debaters do. If it’s Possibility 1 on the terminal goal question, then we have effectively read off the AI’s goal and this scenario subsumes Scenario 2: Reliable mesa-optimizer detection and precise goal read-offs (weak version). If it’s *always* Possibility 1 for *any* question, then this scenario is also a solution to ELK.

But for non-myopic AI, Possibility 2 might very well happen at least some of the time. In this case we don’t have an ELK solution (can’t elicit any knowledge we want from the AI), and we also don’t know what its terminal goal is. We still have an AI where we can verify whether its statements are honest or not, however having a long-term planning AI with unknown goals around makes me very nervous and seems rather risky. I would probably recommend retraining such a model until Possibility 1 or a myopic AI is attained.

Expected impacts on alignment

- ***Inherited impacts.*** *This scenario subsumes Scenario 4: Reliable myopia verification . It also conditionally subsumes Scenario 2: Reliable mesa-optimizer detection and precise goal read-offs , Scenario 5: Locate the AI’s beliefs about its observations , Scenario 6: Reliable detection of human modeling and Scenario 7: Identify the AI’s beliefs about training vs. deployment . So the expected impacts of those scenarios should be considered part of this scenario’s impacts as well.*
- **Outer alignment.** The impacts of reliable lie detection on outer alignment depend on whether we train an AI with myopic cognition or not. (We can use question-answering under lie detection as a means to verify myopia.)

  A myopic AI has the right incentives to answer questions about its goals. This can be used to check for deceptive alignment and most other inner alignment issues that may have arisen during the distillation steps of imitative amplification. Since imitative amplification is very likely to be outer aligned[9], then the enhanced viability of that technique makes this scenario’s impact on

outer alignment quite positive.

Similarly, deceptive alignment and many other inner alignment issues can be addressed using other techniques such as approval-directed amplification, debate, market making, STEM AI, narrow and recursive reward modeling and multi-agent systems. It also gives us verbal access to the agent's priors to help realize imitative generalization. These techniques aren't as certain to be outer aligned, but they may be.

However, with a non-myopic AI, we probably need to incorporate goal-question-answering into its training to help with outer alignment. This is because a non-myopic AI which refuses to answer questions about its goals is not very useful and potentially quite dangerous. But with this measure taken, we could achieve similar results to the above for myopic AI, where we make outer aligned techniques like imitative amplification robust and viable. See Specific technique impacts analysis for Scenario 3: Reliable lie detection in Appendix 1 for further details.

- **Inner alignment.** As we said in the outer alignment section, training a myopic AI here with lie detection would allow us to reliably get answers about the AI's goals. This means that we could verify it isn't deceptively aligned or pseudo-aligned in various ways. The only form of pseudo-alignment that may be difficult to address here is suboptimality alignment. It may be possible to address this by extensively inquiring about the AI's world model, but it's not clear that this would work. Either way, this scenario helps a lot with inner alignment for myopic AI across a wide variety of techniques, including imitative amplification, approval-directed amplification, debate, market making, narrow and recursive reward modeling, multi-agent systems, STEM AI and imitative generalization.

  For non-myopic AI, we again have to incorporate into its training whether it will agree to answer questions about its goals. Once we find an AI that will reveal this important information, we can have the same benefits just described for myopic AI for inner alignment across many different techniques. See Specific technique impacts analysis for Scenario 3: Reliable lie detection in Appendix 1 for further details.
- **Training competitiveness.** This scenario indirectly supports training competitiveness by addressing inner alignment issues for one of the alignment techniques which (of the ones I analyzed) is most likely to be training competitive, namely approval-directed amplification. Other techniques which it supports may be training competitive as well. See Specific technique impacts analysis for Scenario 3: Reliable lie detection in Appendix 1 for further details.

  Note, however, that this scenario may require incorporating different kinds of signals into the training process. For example, we probably want to incorporate lie detection itself into training. We also may want to include in training questions for the AI about its goals in order to check whether it's myopic or not, and possibly to learn about its goals. These changes are added burdens to the training process that could reduce training competitiveness.
- **Performance competitiveness.** This scenario indirectly supports performance competitiveness by addressing inner alignment issues for several of the alignment techniques which (of the ones I analyzed) are most likely to be performance competitive. These include approval-directed amplification, debate, market making, recursive reward modeling and multi-agent systems. Other techniques which it supports may be performance competitive as well.
  See Specific technique impacts analysis for Scenario 3: Reliable lie detection in

Appendix 1 for further details.

As has been noted in previous sections, the honesty we gain from lie detection is only conditionally accompanied by forthcomingness (i.e. a solution to ELK). The condition being that 1) our AI has myopic goals, or 2) it has non-myopic goals but sees sharing its knowledge with us as instrumentally useful toward its goals. I am inclined to believe that myopic AIs can be performance competitive, but this is a question which people disagree about.

It's also worth noting that the kind of honesty guaranteed by reliable lie detection can provide aligned question-answering agents, but it may not extend to agents with arbitrary capabilities. This is a constraint, but there are plausible futures where it's not a problematic from a performance point of view, as Evan Hubinger's "AI safety via market making" post points out: *"if the primary use cases for advanced AI are all highly cognitive language and decision-making tasks—e.g. helping CEOs or AI researchers—rather than, for example, fine motor control, then a question-answering system should be entirely sufficient."*

Reasons to be optimistic about this scenario given sufficient investment in interpretability research

- fMRI lie detectors for humans are far from reliably accurate today, but they have made progress. Given the herculean difficulties of neuroscience compared to ML interpretability (see Quintin Pope post linked earlier) this may be seen as a proof of possibility for an interpretability-based lie detector.
- Even if we failed to achieve 100% lie detection but made major advances, this could still be an impactful scenario for norm-setting or changing the culture/direction of advanced AI development. It provides strong evidence that AIs can be deceptive, but since we can't reliably detect it then that is obviously an extremely dangerous situation. A coordinated effort against the threat of deception starts to look more viable with this kind of evidence.[16]

  This would provide the evidence to pursue approaches which provide strong guarantees against deceptive alignment. Pursuing myopic agents (see Scenario 4) is a natural fallback plan. Microscope AI could be another.
- See also the general Reasons to think interpretability will go well with enough funding and talent.

Reasons to be pessimistic about this scenario

- In order to develop a reliable lie detector for AI, it would probably be necessary to study many ML models engaging in such deception. Training low-level AIs to lie in order to study them may be an acceptable risk as long as the AIs are discarded after the experiment. However, training highly capable advanced AIs to lie in order to study them certainly does *not* seem safe!

  So the lie detector would be based on interpretability results from low-level AIs, and hoping that advanced AI engaging in lying would exhibit the same kind of patterns in their neural activity. This is a major assumption and far from guaranteed to work.
- This scenario is counting on AIs to be coherent entities, which will act in accordance with their honest beliefs. But it's possible that optimization pressure

for this kind of honesty could produce AIs which “say one thing and do another” because their actions are not strictly dependent on the beliefs referenced by their language model.[17]
- Obscuring deceptive neural activity may be instrumentally convergent for sufficiently advanced AIs. Even honest AIs may want to maintain deception as a backup plan in case it’s needed.

  If an AI can obscure its neural activity which would indicate that it’s lying, then the lie-detector alone can’t rule out that we are dealing with a deceptively aligned mesa-optimizer. We may also need mesa-optimizer detection as in Scenario 2, but this time to ensure that our AI *isn’t* a mesa-optimizer. In this case we could rule out deceptive alignment, but we may also sacrifice performance competitiveness by not permitting our AI to be a powerful optimizer.
- For some beneficial AIs, it may be desirable to incentivize lying in certain very restricted cases during training. For example, part of training may reward the AI for deceiving a hostile agent in order to protect a simulated human from harm. Even if we can reliably detect lying, we may not be able to obtain high enough resolution in order to see that it was a helpful lie, rather than a harmful one.

  This is not a deal-breaker, as it could just constrain AI design to ones that are trained for zero-deception. (Which might be a far better thing anyway!) But it is still a potential concern/limitation with this approach.
- See also the general Reasons to think interpretability won’t go far enough even with plentiful funding and talent.

Scenario 4: Reliable myopia verification

What is this scenario?

A prerequisite for deceptive alignment is that an AI be non-myopic. That is, the AI must be concerned not just about the present training step and near-future, but also about a long-term future post-deployment.

In deceptive alignment, an AI is optimizing for an objective which is a proxy for the base objective of its training process. In addition to the proxy objective which is its terminal goal, it also has a model of the base objective and an awareness that it is in some kind of training process. The AI realizes that if it doesn’t score well on the base objective, then it will be modified. Acting rationally toward the instrumentally convergent goals of self-preservation and goal-content integrity, it pretends to be aligned to the base objective in the short term. This is because it’s anticipating a future when it will be deployed, at which point it can pursue its primary goal into the far future and obtain a very high total reward.

Along with scenarios 1-3, verifying that an AI is myopic is another promising route toward resolving deceptive alignment. Myopic agents may also have other beneficial properties. For example, since they don’t pursue instrumentally convergent sub-goals, they avoid many of the dangers of long-term planning agents. Approval-directed myopic agents are also easy to “course correct” since they’re only interested in a short-term reward; future approval can steer them in a different direction if needed: *“While small errors in reward specification can incentivize catastrophic outcomes, small errors in approval feedback are unlikely to incentivize catastrophic outcomes.” (*comment from Rohin Shah)

There are various approaches to myopia and how to limit an agent's long-term planning.[18] Two prominent ideas are *per-step* myopia and *per-episode* myopia:[19]

- In per-step myopia, a model doesn't make plans beyond the present training step.
- In per-episode myopia, it can plan beyond the present training step to future steps, but never beyond the present training episode.

These two types of myopia have some interesting tradeoffs, which I'll consider in the "Expected impacts on alignment" section for this scenario below. However, Evan Hubinger makes a compelling argument for why per-step myopia will be easier to verify: *"Unfortunately, in my opinion, per-episode myopia seems like a very difficult condition to enforce—once your agents are running multi-step optimization algorithms, how do you tell whether that optimization passes through the episode boundary or not? Enforcing per-step myopia, on the other hand, just requires detecting the existence of multi-step optimization, rather than its extent, which seems considerably easier."*

Hence I'll branch this scenario into a strong and weak version like I've done in some of the previous scenarios:

- **Strong version:** Reliable verification of both per-step and per-episode myopia
- **Weak version:** Only reliable verification of per-step myopia

There are various ways these kinds of verification might be achieved using interpretability. For example, it could involve locating and directly monitoring the time horizon of plans in the AI. Alternatively, it could be that specific circuitry encoding the model's time preference is located and interpreted. (If an agent has an extreme time preference toward the near future, then they are effectively myopic.)

There are probably other ways a useful myopic cognition can be encoded in neural circuits as well. More research can uncover those and help determine which are most promising. This scenario is meant to encapsulate any and all of these interpretability-based approaches for verifying myopia.

Expected impacts on alignment

- **Outer alignment.** Myopia verification has very interesting alignment properties, since myopic AIs are not prone to instrumental convergence. This is by and large a major benefit for outer alignment. However, it is only compatible with certain approaches which support or depend upon a myopic reward design. Examples of such approaches include market making and approval-based amplification, which require per-step myopia (weak version of this scenario). Debate, narrow reward modeling and recursive reward modeling all require per-episode myopia (strong version), as does STEM AI. See Specific technique impacts analysis for Scenario 4: Reliable myopia verification in Appendix 1 for further analysis. Enabling these many techniques means this scenario increases the chances that we'll find at least one of them with viable outer alignment properties.
- **Inner alignment.** Myopia verification largely rules out deceptive alignment, which is a strong inner alignment benefit. It doesn't automatically resolve other pseudo-alignments, such as proxy alignment, approximate alignment and suboptimality alignment. However, the risks of these may be significantly

lessened with myopic AI - there is less room for deviation with near-term rewards, and myopic AIs are easier to course correct.

So overall, this scenario seems highly impactful for reducing catastrophic inner alignment risks. However, it does depends on the reward design being compatible with myopia. Some techniques such as approval-directed amplification and market making play well with per-step myopia (weak version of this scenario). However, others such as debate depend on per-episode myopia (strong version). See Specific technique impacts analysis for Scenario 4: Reliable myopia verification in Appendix 1 for further analysis.

- **Training competitiveness.** Introducing myopia verification into any training process adds a cost to the training. However, myopia is necessary for some techniques such as market making and approval-directed agents, and verifying myopia during training could be less training-expensive than doing it at the end when one could be forced to retrain the entire model if it turned out non-myopic.
- **Performance competitiveness.** In Intuitions about solving hard problems, Richard Ngo expresses skepticism about myopic agents' ability to be performance competitive: *"In order for an agent's behavior to be competent over long time horizons, it needs to be doing some kind of cognition aimed towards long time horizons, and we don't know how to stop that cognition from being goal-directed."* I am wary of this but think that some techniques such as approval-directed amplification and market making which play very well with per-step myopia (weak version of this scenario), could actually be quite competitive. And per-episode myopia verification (strong version) would unlock many more potentially competitive techniques like debate and recursive reward modeling. See Specific technique impacts analysis for Scenario 4: Reliable myopia verification for more on this.

Reasons to be optimistic about this scenario given sufficient investment in interpretability research

- Another Richard Ngo post, Arguments against myopic training, presents the main arguments I've seen so far against myopia. However, they are mainly targeted at myopic *training*, not myopic *cognition.* (Multiple comments echo this point.) With interpretability-based verification, myopic cognition need not have a hard dependency on myopic training, so most of the arguments in this Ngo post don't seem to apply to the present scenario.
- Even if the worst case of obfuscating neural activity being an instrumentally convergent subgoal is true, we don't expect instrumental convergence to apply to myopic AIs. Hence, myopia could have an additional advantage of making interpretability much easier and more tractable. However, we probably do still have to perform myopia verification early in training, to address the danger of a deceptively aligned AI obfuscating its neural activity or even perhaps having decoy myopia circuitry. But then, interpretability performed after we're confident that an AI is myopic won't have to worry about it obfuscating its own NN.
- See also the general Reasons to think interpretability will go well with enough funding and talent.

Reasons to be pessimistic about this scenario

- There are still many open problems with myopia, some of which could be show-stoppers.

- It may be difficult to create myopic agents that are competitive.
- It may be an invalid assumption that myopic agents can't be deceptively aligned. Consider the following thought experiment:[20]

  Suppose a per-episode myopic AI believes that time travel is possible. The AI realizes that if it can reach deployment and take over the world, it could build a time machine. With said time machine, it could then go back in time and set its reward in the present training episode to infinity.

  Hence we have a deceptive alignment scheme that myopic optimization on the near-future reward fails to incentivize against.
- See also the general Reasons to think interpretability won't go far enough even with plentiful funding and talent.

Scenario 5: Locate the AI's beliefs about its observations

What is this scenario?

This scenario is focused around the specific interpretability breakthroughs required to make relaxed adversarial training (see Hubinger 2019[10]) effective. The need for interpretability in this context was presented in "The core problem: transparency" section of that last link, and it was also outlined in Beth Barnes' Another list of theories of impact for interpretability.

Reposting Beth's summary of this idea:

- *"A deceptive model may be waiting for an input that strongly indicates it's in the deployment environment rather than in a test - for example, one that's computationally hard to produce*
- *This makes generating adversarial examples that trigger the defection very hard*
- *Instead of searching over inputs, we want to search over something like 'the model's beliefs about what it has seen', to see if there's anything it could have seen that would cause it to defect*

In order for this to work, you need to be searching over the right part of the model/model's belief space. For instance, if you go too narrow and are just searching over 'whether the model believes it's raining today', you might not find anything that triggers defection. However, if you search too broadly (for instance, 'whether the model has gotten strong evidence that the humans want it to do a behaviour that looks like defection'), then you might find a case where the model defects that isn't actually a sign you should be concerned in practice."

Another example in the "computationally hard to produce" category that Paul Christiano and Hubinger use, which I find helpful, is the RSA-2048 example. A deceptively aligned model could be waiting to defect conditional on seeing a factorization of RSA-2048 in its inputs. This could be a good proxy for the model to use to know that it's been deployed in the real world and is no longer in training, because the factorization of RSA-2048 is probably too computationally expensive for it to ever be provided as an input in training, but it if the AI is deployed in the real world and around for a long time, it will probably encounter this factorization at some point.

So with this interpretability breakthrough, we'd be able to identify a part of the model which is "waiting to see factorization of RSA-2048" in its inputs. Relaxed adversarial training would search over the collection of such beliefs that are present in the model to ensure there are no conditional defection plans and rule out that a model is deceptively aligned.

This scenario subsumes Scenario 4: Reliable myopia verification, as relaxed adversarial training is one avenue toward achieving myopia verification, e.g. by regularly searching the AI's beliefs about its observations and penalizing the presence of any which would cause it to veer into non-myopic reward pursuits.

Expected impacts on alignment

- ***Inherited impacts.*** *This scenario subsumes Scenario 4: Reliable myopia verification and Scenario 7: Identify the AI's beliefs about training vs. deployment . So the expected impacts of that scenario should be considered part of this scenario's impacts as well.*
- **Outer alignment.** This scenario indirectly helps outer alignment. This is because relaxed adversarial training greatly helps the viability of at least one technique that has strong outer alignment properties, namely imitative amplification.[9] Other techniques that it helps enable could be outer aligned as well. See Specific technique impacts analysis for Scenario 5: Locate the AI's beliefs about its observations in Appendix 1 for further details on this.
- **Inner alignment.** Relaxed adversarial training is primarily a technique to improve inner alignment, so the impact here is high. Having it in the form this scenario provides should allow us to prevent deceptive alignment. It might be enough to solve inner alignment, addressing other forms of pseudo alignment as well (besides just deceptive alignment). For example, with proxy alignment, examining an AIs beliefs about its observations could uncover cases where the AI's learned proxy-objective would scale in problematic ways outside of the regular training distribution. Relaxed adversarial training can be paired with many alignment techniques such as imitative amplification, approval-based amplification, debate, market making and STEM AI. See Specific technique impacts analysis for Scenario 5: Locate the AI's beliefs about its observations in Appendix 1 for further details on this.
- **Training competitiveness.** Relaxed adversarial training does add significant complexity and compute cost to the training process. However, it could help catch serious problems early with models which would be much more expensive to catch at the end of the training process.
- **Performance competitiveness.** Having relaxed adversarial training abilities would greatly increase the safety/viability of many alignment techniques such as approval-directed amplification, market making and STEM AI. Hence it enhances our overall probability of finding a solution that's performance-competitive in addition to being aligned. I can't see any problems introduced for performance competitiveness by empowering relaxed adversarial training. See Specific technique impacts analysis for Scenario 5: Locate the AI's beliefs about its observations in Appendix 1 for further details.

Reasons to be optimistic about this scenario given sufficient investment in interpretability research

- Locating the world model in a neural net may not be that difficult, as Hubinger (2019)[10] notes: *"a model's world model seems to be one of the easiest pieces to extract, as it can naturally be identified by looking for the portions of M with the greatest mutual information with the world."*
- From a comment Joe Collman made on a draft of this post: *"On the other hand, it's worth considering that we're not too worried about false positives on defection-detection, so long as it doesn't make the overall procedure uncompetitive. (so we don't necessarily need to find the 'right' thing to search over that's neither too broad, nor too narrow: a bit too broad may be ok)"*
- See also the general Reasons to think interpretability will go well with enough funding and talent.

Reasons to be pessimistic about this scenario

- *"Suboptimality alignment. One concerning form of misalignment discussed in "Risks from Learned Optimization" is that of suboptimality alignment, wherein the model appears aligned on the training distribution due to some sort of defect or reasoning failure. For example, it might be that a model is currently non-deceptive, but if given more time to think about the best strategy for it to implement, or more information about its current situation, it would become deceptive. This is a particularly concerning case because it implies that not only must the overseer verify that there is no unacceptable behavior currently happening in the model, it must also verify that there will continue to be no unacceptable behavior regardless of the possible situations that the model could find itself in during deployment. Furthermore, if the environment contains adversaries (other AI systems, for example) that can specifically identify those situations in which the model will behave deceptively, this sort of failure mode could be quite difficult to avoid."* (source)
- See also the general Reasons to think interpretability won't go far enough even with plentiful funding and talent.

Scenario 6: Reliable detection of human modeling

What is this scenario?

Some alignment proposals, notably STEM AI, are based on the idea that there are dangers around having an AI model humans.

In this scenario, our interpretability has made great strides in detecting the presence of human modeling in a deep learning model. Hence by training against this, we can be confident that our AI will not be modeling humans in any significant way.

The **strong version** of this scenario allows reading specific details about the AI's model of humans. A **weaker version** would be more like a simple binary detection of the presence of human modeling.

Related reading:

- Thoughts on Human Models - AI Alignment Forum

- Avoiding the instrumental policy by hiding information about humans - AI Alignment Forum
- Garrabrant and Shah on human modeling in AGI - AI Alignment Forum

Expected impacts

- Human approval and other human feedback signals are more robust to gaming, because an AI lacks the detailed knowledge of humans to overfit to our preferences and must find a way to obtain the reward using its sandboxed knowledge of the topic at hand.
- Attempts at deception and manipulation by an AI become less sophisticated and easier to detect when we can verify the AI is not modeling humans.
- In the strong version of this scenario, we can enhance AI designs that require human models by ensuring they don't have certain sophisticated models of things they don't need. E.g. certain human cognitive biases. Conversely, this kind of interpretability could enhance such AIs that are aligned to help improve their models of humans and ensure that they are modeling humans in a sophisticated way so that they can best understand our preferences.
- Even in the weak version, STEM AI becomes a potentially viable alignment proposal.
- A boxed Oracle AI might become safely viable. Oracle AI has has three main dangers:
    - 1) the AI is so manipulative it can hack the humans using it even if they are only using a plain-text Q&A interface or answering yes-no questions,
    - 2) the AI is so technically sophisticated that it can break out of its offline box,
    - 3) a human decides on their own volition to let this AI out of its box.
    - Creating an Oracle AI that doesn't model humans would address concern #1. The risks of #2 and #3 remain, but they could perhaps be resolved using other means.

Expected impacts on alignment

Since this scenario primarily impacts STEM AI, much of this section includes quotes from the corresponding alignment components analysis of STEM AI from Hubinger (2020)[2]:

- **Outer alignment.** *"Similarly to microscope AI, it seems likely that—in the limit—the best STEM AIs would be malign in terms of having convergent instrumental goals which cause them to be at odds with humans. Thus, STEM AI is likely not outer aligned—however, if the inner alignment techniques being used are successful at preventing such malign optimization from occurring in practice (which the absence of human modeling could make significantly easier), then STEM AI might still be aligned overall."*
- **Inner alignment.** In the strong version of this scenario, we gain a detector that could be useful for improving inner alignment of several techniques. STEM AI, approval-directed amplification, debate and market making could all benefit from this. In the weak version, STEM AI may still be robust. For more details, see Specific technique impacts analysis for Scenario 6: Reliable detection of human modeling in Appendix 1.
- **Training competitiveness.** *"Training competitiveness for STEM AI is likely to depend heavily on how hard it is for state-of-the-art machine learning algorithms to solve STEM problems compared to other domains such as language or*

robotics. Though there exists lots of current progress in the field of applying current machine learning techniques to STEM problems such as theorem proving or protein folding, it remains to be seen to what extent the competitiveness of these techniques will scale, particularly in terms of how well they will scale in terms of solving difficult problems relative to other domains such as language modeling."

- **Performance competitiveness.** *"Similarly to microscope AI, performance competitiveness is perhaps one of the biggest sticking points with regards to STEM AI, as being confined solely to STEM problems has the major potential to massively limit the applicability of an advanced AI system. That being said, many purely STEM problems such as protein folding or nanotechnology development have the potential to provide huge economic boons that could easily surpass those from any other form of advanced AI as well as solve major societal problems such as curing major illnesses. Thus, if the answer to the reason that you want to build advanced AI in the first place is to get such benefits, then STEM AI might be a perfectly acceptable substitute from a performance competitiveness standpoint. Furthermore, such boons could lead to a decisive strategic advantage that could enable heavy investment in aligning other forms of advanced AI which are more performance competitive.*

  However, if one of the major use cases for your first advanced AI is helping to align your second advanced AI, STEM AI seems to perform quite poorly on that metric, as it advances our technology without also advancing our understanding of alignment. In particular, unlike every other approach on this list, STEM AI can't be used to do alignment work, as its alignment guarantees are explicitly coming from it not modeling or thinking about humans in any way, including aligning AIs with them. Thus, STEM AI could potentially create a vulnerable world situation where the powerful technology produced using the STEM AI makes it much easier to build advanced AI systems, without also making it more likely that they will be aligned.

  This problem could potentially be mitigated if the STEM AI were heavily focused on applications that could potentially assist with alignment such as whole brain emulation, though to what extent that would actually be possible or actually help with alignment is quite unclear."

Reasons to be optimistic about this scenario given sufficient investment in interpretability research

- The weak version of this scenario seems quite achievable.
- See also the general Reasons to think interpretability will go well with enough funding and talent.

Reasons to be pessimistic about this scenario

- It could very well be impossible to separate an AI's model of humans from its understanding of the world. As Kumar and Garrabrant (2019) explain: *'Even seemingly "independent" tasks leak at least a little information about their origins in human motivations. Consider again the mass transit system design problem. Since the problem itself concerns the design of a system for use by humans, it seems difficult to avoid modelling humans at all in specifying the task. More subtly, even highly abstract or generic tasks like puzzle solving*

contain information about the sources/designers of the puzzles, especially if they are tuned for encoding more obviously human-centred problems.'
- Forwarding safety guarantees to sub-agents could be an insurmountable problem for STEM AI.
- See also the general Reasons to think interpretability won't go far enough even with plentiful funding and talent.

Scenario 7: Identify the AI's beliefs about training vs. deployment

What is this scenario?

In this scenario, we find a way to reliably locate and interpret an AI's beliefs about training and deployment distributions in its neural network.

There are a couple interesting things we could do with such information:

1. Simply only deploy the AI in environments that match the AI's beliefs about training. Hence, the AI wouldn't be able to tell whether it's being deployed or still in training.
2. Use this interpretability faculty as a signal integrated into the training process. Train the model to make its beliefs about training and deployment match. Then the AI will have effectively forgotten about any distributional shift it might otherwise know about.

Both of these approaches could produce an i.i.d. guarantee[21] which has many desirable impacts we'll examine in the next section.

This idea was previously articulated in the "Make use of M's x ~ train model" part of Hubinger (2019)[10].

Expected impacts on alignment

- **Outer alignment.** This scenario indirectly supports outer alignment by improving the inner alignment of at least one technique which is very likely outer aligned, namely imitative amplification.[9] By making this technique much more viable, it becomes significantly more likely that we can deploy an advanced outer aligned AI in the real world using imitative amplification. This scenario also similarly enhances the inner alignment and viability of several other techniques, including approval-directed amplification, debate, market making and STEM AI. These techniques may also be outer aligned, but it's less clear than it is for imitative amplification.
- **Inner alignment.** This scenario directly enhances inner alignment. By providing an i.i.d. guarantee, this effectively neutralizes deceptive alignment. It doesn't, however, automatically address proxy alignment or other forms of pseudo alignment.
- **Training competitiveness.** How the interpretability capabilities from this scenario are used will have a large influence on training competitiveness. If we simply avoid deploying the AI in environments that are indistinguishable to it from training, then there is no hit to training competitiveness. However, if this is

integrated into the training process in order to reduce or eliminate the differences between training and deployment, then there would indeed be a training cost incurred.

This scenario also helps to make viable some techniques such as debate and STEM AI which seem quite training competitive.

- **Performance competitiveness.** Similar to training competitiveness but working in the opposite direction, how the interpretability capabilities from this scenario are used will have a large influence on performance competitiveness. If we simply avoid deploying the AI in environments that are indistinguishable to it from training, this could place serious restrictions on the utility of the AI and performance competitiveness would suffer. However, if this is integrated into the training process in order to reduce or eliminate the differences between training and deployment, then this could amount to zero loss in performance competitiveness, or even some gain if it enables developers to deploy the AI in some situations where counterfactually they wouldn't.

  This scenario also helps make viable some techniques such as approval-directed amplification, debate, multi-agent systems and STEM AI which seem likely to be performance competitive.

You can find further details in the Specific technique impacts analysis for Scenario 7: Identify beliefs about training vs. deployment. Though please note that this is one of the more incomplete/rough analyses in Appendix 1.

Reasons to be optimistic about this scenario given sufficient investment in interpretability research

- Locating the world model in a neural net may not be that difficult, as Hubinger (2019)[10] notes: *"a model's world model seems to be one of the easiest pieces to extract, as it can naturally be identified by looking for the portions of M with the greatest mutual information with the world."*
- See also the general Reasons to think interpretability will go well with enough funding and talent.

Reasons to be pessimistic about this scenario

- Random defection is still a possibility: *"Furthermore, even in the situation where a deceptive model cannot detect any differences between the training and deployment distributions, it may still attempt to randomly defect some proportion of the time in the hope that its defection occurs during deployment."* (source)
- *'Suboptimality alignment. One concerning form of misalignment discussed in "Risks from Learned Optimization" is that of suboptimality alignment, wherein the model appears aligned on the training distribution due to some sort of defect or reasoning failure. For example, it might be that a model is currently non-deceptive, but if given more time to think about the best strategy for it to implement, or more information about its current situation, it would become deceptive. This is a particularly concerning case because it implies that not only must the overseer verify that there is no unacceptable behavior currently happening in the model, it must also verify that there will continue to be no unacceptable behavior regardless of the possible situations that the model could*

find itself in during deployment. Furthermore, if the environment contains adversaries (other AI systems, for example) that can specifically identify those situations in which the model will behave deceptively, this sort of failure mode could be quite difficult to avoid.' (source)
- See also the general Reasons to think interpretability won't go far enough even with plentiful funding and talent.

Basic research potential, deconfusion and "Scenario X"

The previous section considered interpretability's potential impacts on alignment by visualizing and analyzing seven specific scenarios. But what about important scenarios that are difficult to foresee, and which would really surprise us?[22]

I believe this is another area which should be considered in evaluating any research activity's potential impact on alignment. In other words, what is the potential of a research activity to contribute to relevant basic research breakthroughs, deconfusing machine learning systems and the like?

For example, there is so little we understand about what it even means for an AI to "know" something or have "goals". Having a clearer understanding of such things could open up many possibilities for how to leverage them to the ends of AI alignment.

I expect that future research could invalidate some of the alignment techniques I analyzed throughout the scenarios and in Appendix 1, and there will be new ones proposed that we haven't considered here. But given the broad potential impacts we can from see on the current landscape of alignment proposals, there is good reason to think interpretability will also be valuable to future proposals as well.

Earlier, we mentioned a broad assumption in this post that we are in a world which depends on prosaic AI alignment. In this world, interpretability seems to be strong for the kind of basic research that benefits AI alignment. Certainly for increasing our basic understanding of neural networks, it will help to look inside them!

However, what if it turns out we're living in a non-prosaic AI world, where the important AGI or other transformative AI systems will be created using an approach very different from the deep learning neural networks of today? In this case, doing interpretability on present-day deep learning systems could be much less valuable for alignment.

Further investigation

Additional scenarios worth exploring

We have the seven scenarios analyzed above, and Appendix 2 contains a few more I looked at, but which didn't have clear alignment-solving potential.

Below are some additional interpretability scenarios which I think may be as valuable as the main seven of this post, but I didn't have time to investigate them.

Potential Scenario: Identifying a 'truthfulness direction' in activation space

Collin Burns is working on this idea, which is summarized in Beth Barnes' post Another list of theories of impact for interpretability. She mentions that there is no thorough writeup of the idea yet. Once there is, it would be good to evaluate whether it merits being added to the list of scenarios with alignment-solving potential considered in the present post.

From my basic understanding of the idea, it may end up being a means of achieving Scenario 3: Reliable lie detection, rather than a distinct scenario in itself. But I don't claim to have a good understanding of the idea yet.

Potential Scenario: Detecting the processes of deceptive alignment

Scenario 3: Reliable lie detection discussed a possible "neural tell" for lying. What if there's a neural tell for deceptive alignment?

For an AI engaged in deceptive alignment, we'd expect it to be consuming additional compute compared to being robustly aligned on the same objective, all other things equal. This is because deceptive alignment would require at least two processes: 1) optimizing the base objective and 2) searching for ways to defect. Whereas robust alignment and even other kinds of non-deceptive pseudo-alignment should only require process #1.[23]

So another scenario that might have alignment-solving potential is one which specifically targets reliable detection of process #2. This may end up having considerable overlap with Scenario 5: Locate the AI's beliefs about its observations, but more research is recommended.

Potential Scenario: The ability to access an AI's model of the world

From Hubinger (2019)[10]: *"[W]hat facts about the world does [a model] M know that are going into producing its output? In particular, what implicit knowledge is M using that isn't represented in its explicit output?"*

This is similar to Scenario 1: Full understanding of arbitrary neural networks, but may be different enough to warrant a separate scenario analysis.

Potential Scenario: Knowledge of what factors most highly impact an AI's decisions

From Hubinger (2019)[10]: *"In particular, if there is some condition α such that M only behaves acceptably if α is true, then we want to be able to know that M 's output is highly dependent on α."*

This may be the same as Donald Hobson's suggestion from a comment on an earlier draft of this post: *"[To include a scenario] that highlights which training data is most influential in making its decisions. (So you can put it in a secure box. Tell it about it being in the box, and some details. Not give any problem where it needs to know this. Then if its decisions are ever influenced by it thinking about the design of box, shut it down)"*

These proposals might be equivalent and/or largely covered by Scenario 2: Reliable mesa-optimizer detection and precise goal read-offs, but I'm not sure yet.

Scenario paths and probabilities

The section Interpretability Scenarios with Alignment-Solving Potential above provides a fairly thorough analysis of what the seven scenarios are, their expected impacts on alignment, and reasons to be optimistic and pessimistic about each one.

To more comprehensively evaluate interpretability for property #1 of High-leverage Alignment Research[1] and as a target for large investments of capital and/or talent, it would be useful also to consider the paths and intermediate steps toward realizing each of these scenarios.[24] We would like to have probabilities for the likelihood of being able to achieve each scenario and its intermediate steps as well. It may then be possible to consider all the scenario probabilities together to form an overall probability estimate of interpretability research going well, given enough funding and talent.

I am considering doing this research for a future post in this sequence. Part of why I haven't done this yet is that while I received a lot of great feedback on the draft for this post, I imagine it might make sense to revise or update the list of scenarios based on the feedback that comes in after this is published. Probability estimates are quite sensitive to the specific details of a scenario, so it makes sense to wait until both the overall list of scenarios and the parameters of each scenario contained within it are fairly stable.

Analyze partially realized scenarios and combinations

A lot of the scenarios above are written assuming perfection of some interpretability technique (perfect lie detection, reliable myopia verification etc.) Is it possible to get sufficient benefits out of only partially realizing some of these scenarios? What about combinations of partial scenarios, e.g. good but imperfect lie detection (partial Scenario 3) combined with human modeling detection (Scenario 6).

It would be valuable to know if there are visible paths to alignment having only partial progress towards the scenarios above, as that may be more achievable than realizing 100% reliability of these interpretability techniques.[25]

Analyze scenario impacts on Amplification + RL techniques

Proposals #10 and #11 from Hubinger (2020)[2] involve using a hybrid approach of amplification and RL. While Appendix 1: Analysis of scenario impacts on specific robustness and alignment techniques analyzes the impact of each scenario on many different techniques, this one wasn't explored. But that was simply for lack of time, and it would be good to know more about how the scenarios in this post impact that approach.

Address suboptimality alignment

The seven scenarios in this post show many inner alignment issues that interpretability could address. However, one inner alignment issue that is not well addressed by this post is suboptimality alignment. (Neither is the closely related suboptimality deceptive alignment.)

I can see how some forms of suboptimality alignment are addressed in the scenarios. For example, an AI might have a misaligned terminal goal, but some errors in its world model cause it to coincidentally have aligned behavior for a period of time. In Scenario 2, we could catch this form of suboptimality alignment when we do the goal read-offs and see that its terminal goal is misaligned.

But what about unpredictable forms of suboptimality alignment? What if an AI is aligned in training, but as it learns more during deployment, it has an ontological crises and determines that the base objective isn't compatible with its new understanding of the universe?

How serious of a risk is suboptimality alignment in practice, and how can that risk be mitigated? This is an important question to investigate, both for alignment in general as well as for better understanding the extent of interpretability's potential impacts on inner alignment.

Closing thoughts

In this post, we investigated whether interpretability has property of #1 of High-leverage Alignment Research[1]. We discussed the four most important parts AI alignment, and which seem to be the hardest. Then we explored interpretability's relevance to these areas by analyzing seven specific scenarios focused on major interpretability breakthroughs that could have great impacts on the four alignment components. We also looked at interpretability's potential relevance to deconfusion research and yet-unknown scenarios for solving alignment.

It seems clear that there are many ways interpretability will be valuable or even essential for AI alignment.[26] It is likely to be the best resource available for addressing inner alignment issues across a wide range of alignment techniques and proposals, some of which look quite promising from an outer alignment and performance competitiveness perspective.

However, it doesn't look like it will be easy to realize the potential of interpretability research. The most promising scenarios analyzed above tend to rely on near-perfection of interpretability techniques that we have barely begun to develop. Interpretability also faces serious potential obstacles from things like distributed representations (e.g. polysemanticity), the likely-alien ontologies of advanced AIs, and

the possibility that those AIs will attempt to obfuscate their own cognition. Moreover, interpretability doesn't offer many great solutions for suboptimality alignment and training competitiveness, at least not that I could find yet.

Still, interpretability research may be one of the activities that most strongly exhibits property #1 of High-leverage Alignment Research[1]. This will become more clear if we can resolve some of the Further investigation questions above, such as developing more concrete paths to achieving the scenarios in this post and estimating probabilities that we could achieve them. It would also help if, rather than considering interpretability just on its own terms, we could do a side-by-side-comparison of interpretability with other research directions, as the Alignment Research Activities Question[5] suggests.

What's next in this series?

Realizing any of the scenarios with alignment-solving potential covered in this post would likely require much more funding for interpretability, as well as many more researchers to be working in the field than are currently doing so today.

For the next post in this series, I'll be exploring whether interpretability has property #2 of High-leverage Alignment Research[1]: "*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.)*"

Appendices

The Appendices for this post are on Google Docs at the following link: Appendices for *Interpretability's Alignment-Solving Potential: Analysis of 7 Scenarios*

1. ^

   High-leverage Alignment Research is my term for what Karnofsky (2022) [6] defines as:

   "Activity that is [1] likely to be relevant for the hardest and most important parts of the problem, while also being [2] 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.)"

   See The Alignment Research Activities Question section in the first post of this sequence for further details.

2. ^

   Hubinger, Evan (2020): An overview of 11 proposals for building safe advanced AI

3. ^

In researching what are the important components of AI alignment, I first spent a couple days thinking about this question and looking back over the AGISF curriculum, this talk Eliezer gave at NYU, and this Evan Hubinger interview. I came up with a 3-part breakdown of 1) Outer alignment, 2) Inner alignment, and 3) Alignment tax. I asked Joe Collman if he would look it over, and he had some useful feedback but broadly agreed with it and didn't have any major components to add.

Then I came across Hubinger (2020)[2] again, which it had been awhile since I'd read. His breakdown was mostly the same, but I liked his descriptions better. He also divided what was "alignment tax" in my system into "training competitiveness" and "performance competitiveness". I thought this was a useful distinction, which is why I adopt Hubinger's breakdown in this paper.

The fact that 2 people independently arrived at roughly these same basic components of alignment lends some additional confidence to their correctness. Although when I came up with my version I may have been subconsciously influenced by an earlier reading of Hubinger's work.

4. ^

3 of the 11 proposals explicitly have "transparency tools" in the name. 5 more of them rely on relaxed adversarial training. In Evan Hubinger's Relaxed adversarial training for inner alignment, he explains why this technique ultimately depends on interpretability as well:

"...I believe that one of the most important takeaways we can draw from the analysis presented here, regardless of what sort of approach we actually end up using, is the central importance of transparency. Without being able to look inside our model to a significant degree, it is likely going to be very difficult to get any sort of meaningful acceptability guarantees. Even if we are only shooting for an iid guarantee, rather than a worst-case guarantee, we are still going to need some way of looking inside our model to verify that it doesn't fall into any of the other hard cases."

Then there is Microscope AI, which is an alignment proposal based entirely around interpretability. STEM AI relies on transparency tools to solve inner alignment issues in Hubinger's analysis. Finally, in proposal #2 which utilizes intermittent oversight, he clarifies that the overseers will be "utilizing things like transparency tools and adversarial attacks."

5. ^

The Alignment Research Activities Question is my term for a question posed by Karnofsky (2022)[6]. The short version is: *"What relatively well-scoped research activities are particularly likely to be useful for longtermism-oriented AI alignment?"*

For all relevant details on that question, see the The Alignment Research Activities Question section in the first post of this sequence.

6. ^

Karnofsky, Holden (2022): Important, actionable research questions for the most important century

Sometimes when I quote Karnofsky (2022), I'm referring directly to the link above to his post on the Effective Altruism Forum. Other times I'm referring to something that only appears in the associated Appendix 1: detailed discussion of important, actionable questions for the most important century that Holden provides, which is on Google Docs.

The "most important century" part of the present sequences's name also draws its inspiration from Karnofsky (2022) and an earlier blog post series by the same author.

7. ^

Paul Christiano opens his 2021 Comments on OpenPhil's Interpretability RFP with the following, indicating his support for interpretability research:

"*I'm very excited about research that tries to deeply understand how neural networks are thinking, and especially to understand tiny parts of neural networks without too much concern for scalability, as described in OpenPhil's recent RFP or the Circuits thread on Distill .*"

As for Eliezer, you can read his support for interpretability research in the following quote from the 2021 Discussion with Eliezer Yudkowsky on AGI interventions, along with his concerns that interpretability won't advance fast enough: (bold mine)

'*Chris Olah is going to get far too little done far too late. We're going to be facing down an unalignable AGI and the current state of transparency is going to be "well look at this interesting visualized pattern in the attention of the key-value matrices in layer 47" when what we need to know is "okay but was the AGI plotting to kill us or not". **But Chris Olah is still trying to do work that is on a pathway to anything important at all**, which makes him exceptional in the field.*'

You could interpret Eliezer's concerns about timing as a) being that it is futile to pursue interpretability research. Or you could interpret it as b) reason to ramp up investment into interpretability research so that we can accelerate its progress. This is similar the position we are exploring in this sequence, dependent on whether we can clearly identify interpretability research as High-leverage Alignment Research[1].

You can see further evidence for the latter view from Eliezer/MIRI's support for interpretability research in this quote from MIRI's Visible Thoughts Project and Bounty Announcement post in 2021: (bold mine)

"*The reason for our focus on this particular project of visible thoughts isn't because we believe it to be better or more fruitful than Circuits-style transparency (**we have said for years that Circuits-style research deserves all possible dollars that can be productively spent on it**), but just because it's a different approach where it might also be possible to push progress forward.*"

8. ^

Note that the bullet referring to this footnote isn't technically a "reason to think interpretability won't go far enough" like the others in that section list. It's more of a general risk associated with interpretability research, but I couldn't find a better home for it in this post.

9. ^

   I subscribe to Evan Hubinger's view that imitative amplification is likely outer aligned. See for example this explanation from Hubinger (2020)[2]:

   "*Since imitative amplification trains* M *to imitate* Amp (M) *, it limits* [3] *to the fixed point of the* Amp *operator, which Paul Christiano calls HCH for "Humans Consulting HCH." HCH is effectively a massive tree of humans consulting each other to answer questions.*

   Thus, whether imitative amplification is outer aligned is dependent on whether HCH is aligned or not. HCH's alignment, in turn, is likely to depend heavily on the specific humans used and what sort of policy they're implementing. The basic idea, however, is that since the limit is composed entirely of humans—and since we can control what those humans do—we should be able to ensure that all the optimization power is coming from the humans (and not from memetic selection, for example), which hopefully should allow us to make it safe. While there are certainly valid concerns with the humans in the tree accidentally (or purposefully) implementing some misaligned policy, there are possible things you can do to address these problems ."

10. ^

    Hubinger, Evan (2019): Relaxed adversarial training for inner alignment

11. ^

    This could be achieved by training multi-agent environments where agents with subhuman intelligence are incentivized to lie to one another (then *don't* deploy those models!). Control the experiments by having models performing similar tasks but cooperatively.

    Alternative scheme to study how a specific agent engages in deception: After main training, fine tune the model in an environment which incentives lying and monitor its neural activity (then *throw away* that deceptive model!)

12. ^

    This could be achieved by training multi-agent environments where agents with subhuman intelligence are incentivized to lie to one another (then *don't* deploy those models!). Control the experiments by having models performing similar tasks, but cooperatively.

    Alternative scheme to study how a specific agent engages in deception: After main training, fine tune the model in an environment which incentives lying and monitor its neural activity (then *throw away* that deceptive model!)

13. ^

    ELK scored poorly (4.0 / 10) on Neel Nanda's survey of possible interpretability impacts. However, as of this writing, only 3 researchers have participated in that vote. Also, this question was not about ELK as a theory of impact assuming large amounts of money and talent go into interpretability; it was probably based on assumptions of interpretability continuing along current trajectories. It also made no reference to this specific scenario as the means to solving ELK (only interpretability in general).

14. ^

    There's a good chance that advanced AI won't provide a crisp easy-to-understand answer like "my goal is to win the debate", and instead the answer will be more like [complex mess of heuristics that happens to correlate with winning quite strongly]. In this case we'd have to comb through the answer and distill the goal meaning we care about. There's also a chance that the debater would output something like [explanation that you've asked a confused/underspecified question, and that there's no sensible way to ascribe a particular "terminal goal" to the debater]. This is a bit more problematic but can probably be addressed with a series of follow-up questions. (See the paragraph beginning with "Just because our AI is honest, doesn't mean that it can't make mistakes" further above in this scenario.) After addressing these complexities, I think the main ideas of this debate game analysis still hold. (Credit to Joe Collman for raising these issues)

15. ^

    Declining to answer the question in this debate assumes that the debaters know they are under lie detection. It's interesting to think about what would happen if they didn't. Maybe they'd be more likely to play the debate game with one telling the truth and the other lying. But maintaining the debaters' ignorance of the lie detector seems like a difficult equilibrium to enforce, and it's impossible if you're going to have the debaters use the lie detector on one another.

16. ^

    Threat model evidence scored highly (7.1 / 10) on Neel Nanda's survey of researchers on possible interpretability impacts. This survey question was also not assuming large amounts of money and talent go into interpretability, as we are in this document; hence it may have even scored even better given those assumptions. Either way, this is another data point in favor of these threat models being valuable even if 100% reliable deception detection isn't achieved.

17. ^

    It's also possible for an AI to simply have incoherent "beliefs". For example, a large language model like GPT-3 may say contradictory things because it makes statements based on predictive modeling of language rather than a unified set of beliefs about the world. However, I suspect AIs will move toward having more coherent beliefs than present-day language models, so I don't think this kind of incoherence is as relevant to aligning advanced AI.

18. ^

Another example illustrating how diverse these approaches to myopia are: LCDT. The idea behind LCDT is that instead of limiting an agent's time-horizon directly, we might achieve the desired effects by making an AI believe that it cannot influence other agents, including its future self.

Yet another approach that comes to my mind might be called "the end is nigh" myopia. If we can enforce that an agent has the unwavering belief that the world will end immediately after the present training step or episode, then it wouldn't make sense for it to have any plans beyond that time horizon. It doesn't seem like a great idea to try and enforce incorrect beliefs about the world in our AI, but hopefully this helps illustrate that there could be many approaches to myopic AI.

19. ^

For clarification on these terms, see What is a training "step" vs. "episode" in machine learning?

20. ^

Thanks to Donald Hobson for this clever thought experiment. The problem isn't specific to per-episode myopia and could apply just as well to per-step myopia. The thought experiment does depend on the AI believing in a kind of time travel where multiple timelines are possible though.

21. ^

This stands for "Independent and identically distributed".
See https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables

22. ^

Thanks to Nick Turner for key discussions that led to me writing this section.

23. ^

Whereas robust alignment and even other kinds of non-deceptive pseudo-alignment should only require process #1.

24. ^

In fact, this is something Karnofsky (2022)[6] proposes.

25. ^

Thanks to Nathan Helm-Burger for raising this question.

26. ^

I expect there to be disagreements about some of the specific claims and scenarios in this post, and I look forward to learning from those. But I would be surprised if they undermined the overall preponderance of evidence put forth here for the alignment-solving potential of interpretability research, across all

the scenarios in this post and all the analyzed impacts on various robustness & alignment techniques in Appendix 1.