Inaccessible Information

Paul F. Christiano

3rd of June 2020

Suppose that I have a great model for predicting "what will Alice say next?"

I can evaluate and train this model by checking its predictions against reality, but there may be many facts this model "knows" that I can't easily access.

For example, the model might have a detailed representation of Alice's thoughts which it uses to predict what Alice will say, *without* being able to directly answer "What is Alice thinking?" In this case, I can only access that knowledge indirectly, e.g. by asking about what Alice would say in under different conditions.

I'll call information like "What is Alice thinking?" inaccessible. I think it's very plausible that AI systems will build up important inaccessible knowledge, and that this may be a central feature of the AI alignment problem.

In this post I'm going to try to clarify what I mean by "inaccessible information" and the conditions under which it could be a problem. This is intended as clarification and framing rather than a presentation of new ideas, though sections IV, V, and VI do try to make some small steps forward.

I. Defining inaccessible information

I'll start by informally defining what it means for information to be **accessible**, based on two mechanisms:

Mechanism 1: checking directly

If I can check X myself, *given other accessible information,* then I'll define X to be accessible.

For example, I can check a claim about what Alice will do, but I can't check a claim about what Alice is thinking.

If I can run randomized experiments, I can probabilistically check a claim about what Alice *would* do. But I can't check a counterfactual claim for conditions that I can't create in an experiment.

In reality this is a graded notion—some things are easier or harder to check. For the purpose of this post, we can just talk about whether something can be tested even a single time over the course of my training process.

Mechanism 2: transfer

The simplest model that provides some accessible information X may also provide some other information Y. After all, it's unlikely that the simplest model that outputs X doesn't output *anything* else. In this case, we'll define Y to be accessible.

For example, if I train a model to predict what happens over the next minute, hour, or day, it may generalize to predicting what will happen in a month or year. For example, if the simplest model to predict the next day was a fully-accurate physical simulation, then the same physics simulation might work when run for longer periods of time.

I think this kind of transfer is kind of dicey, so I genuinely don't know if long-term predictions are accessible or not (we certainly can't directly check them, so transfer is the only way they could be accessible).

Regardless of whether long-term predictions are accessible by transfer, there are other cases where I think transfer is pretty unlikely. For example, the simplest way to predict Alice's behavior might be to have a good working model for her thoughts. But it seems unlikely that this model would spontaneously describe what Alice is thinking in an understandable way—you'd need to specify some additional machinery, for turning the latent model into useful descriptions.

I think this is going to be a fairly common situation: predicting accessible information may involve almost all the same work as predicting inaccessible information, but you need to combine that work with some "last mile" in order to actually output inaccessible facts.

Definition

I'll say that information is *accessible* if it's in the smallest set of information that is closed under those two mechanisms, and *inaccessible* otherwise.

There are a lot of nuances in that definition, which I'll ignore for now.

Examples

Here are some candidates for accessible vs. inaccessible information:

- "What will Alice say?" vs "What is Alice thinking?"
- "What's on my financial statement?" vs. "How much money do I really have?"
- "Am I coughing?" vs. "What's happening with my immune system?"

- “How will senators vote?” vs. “What’s the state of political alliances and agreements in the senate?”
- “What do I see on my computer screen?” vs. “Is my computer compromised?”
- “What’s the market price of this company?” vs. “How valuable is this IP really?”
- “Will the machine break tomorrow?” vs. “Is there hard-to-observe damage in this component?”
- “What does the news show me from 5000 miles away?” vs. “What’s actually happening 5000 miles away?”
- “Is this argument convincing?” vs. “Is this argument correct?”
- “What will happen tomorrow?” vs. “What will happen in a year” (depending on whether models transfer to long horizons)

II. Where inaccessible info comes from and why it might matter

Our models can build up inaccessible information because it helps them predict accessible information. They know something about what Alice is thinking because it helps explain what Alice does. In this diagram, the black arrow represents the causal relationship:

Unfortunately, this causal relationship doesn’t directly let us *elicit* the inaccessible information.

Scientific theories are prototypical instances of this diagram, e.g. I might infer the existence of electron from observing the behavior of macroscopic objects. There might not be any explanation for a theory other than “it’s made good predictions in the past, so it probably will in the future.” The actual claims the theory makes about the world—e.g. that the Higgs boson has such-and-such a mass—are totally alien to someone who doesn’t know anything about the theory.

I’m not worried about scientific hypotheses in particular, because they are usually *extremely* simple. I’m much more scared of analogous situations that we think of as intuition—if you want to justify your intuition that Alice doesn’t like you, or that some code is going to be hard to maintain, or that one tower of cards is going to be more stable than another, you may not be able to say very much other than “This is part of a complex group of intuitions that I built up over a very long time and which seems to have a good predictive track record.”

At that point “picking the model that matches the data best” starts to look a lot like doing ML, and it’s more plausible that we’re going to start getting hypotheses that we don’t understand or which behave badly.

Why might we care about this?

In some sense, I think this all comes down to what I’ve called strategy-stealing: if AI can be used to compete effectively, can humans use AI to compete *on their*

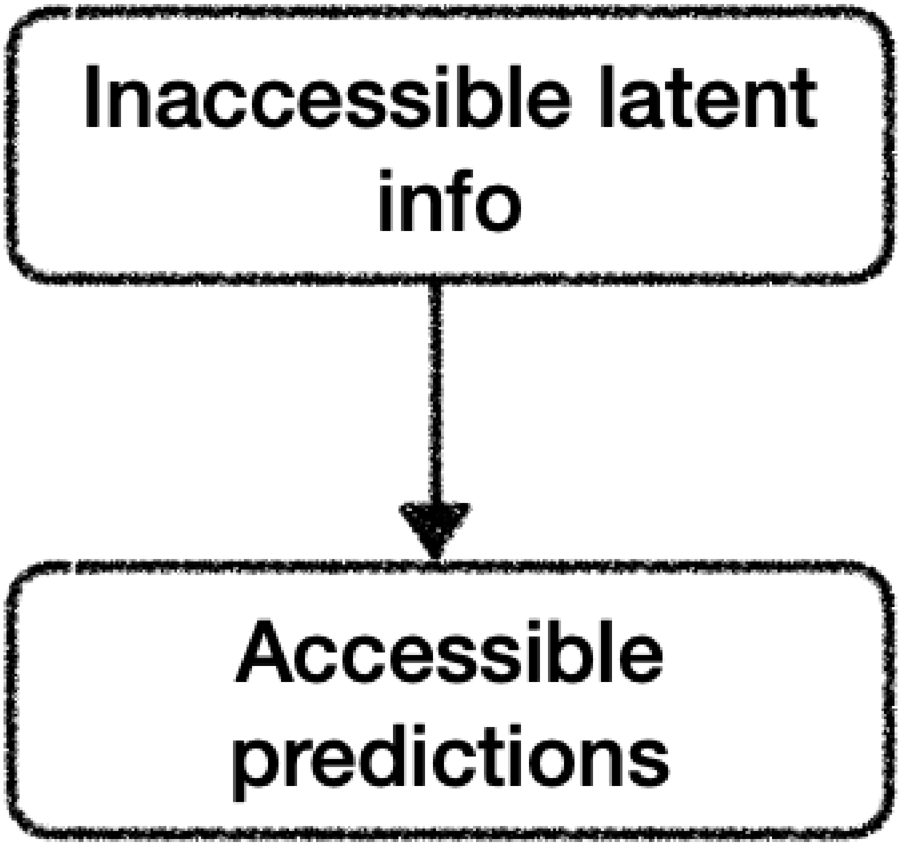

Figure 1:

behalf?

More precisely, for every strategy A that an AI could pursue to bring about some arbitrary outcome, is there a strategy A* that would help humans get what we want over the long term, without leaving us at a competitive disadvantage over the short term?

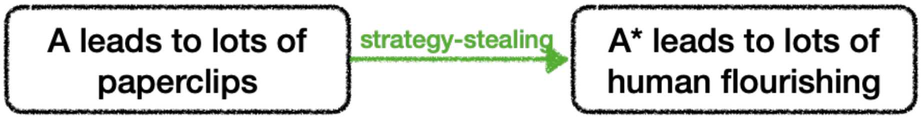
A leads to lots of paperclips
strategy-stealing
A* leads to lots of human flourishing

Figure 2:

If so it's good news for humanity: if most humans build AIs who execute plans like A*, then humans won't be outcompeted by unaligned AIs who execute plans like A.

But the mere *existence* of A* isn't very helpful, we need to actually be able to figure out that A* leads to human flourishing so that we can do it. If we can't recognize plans like A*, then humanity will be at a disadvantage.

We could have a problem if the fact "A* leads to human flourishing" is inaccessible while the fact "A leads to paperclips" is accessible.

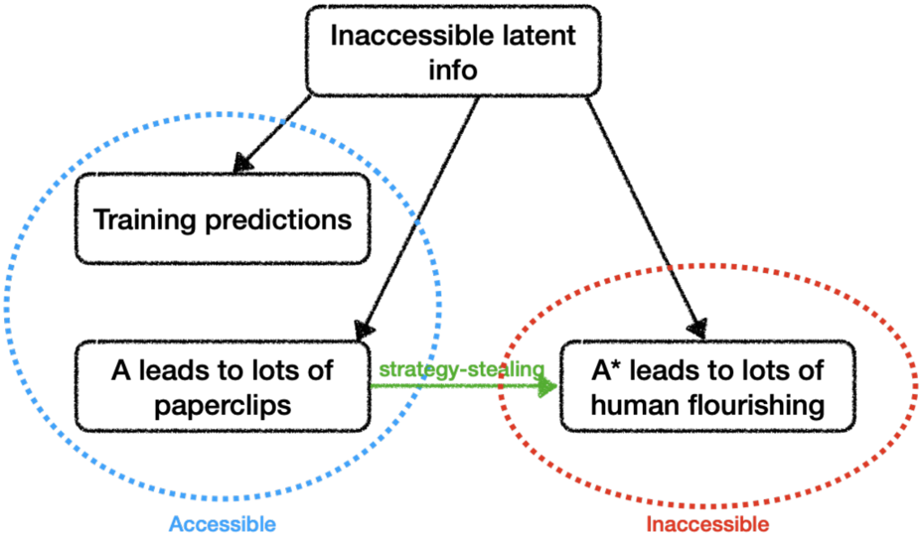

Figure 3:

In this case, we either need to be sure that no one acts on the basis of information like "A leads to lots of paperclips," or we need to figure out some way to access the inaccessible information that "A* leads to lots of human flourishing."

III. Eliciting inaccessible info and the instrumental policy

I've been talking as if we can't access inaccessible info (and the name is suggestive…), but I don't think you should be convinced of that yet.

Let M be a model that reports only accessible facts—it answers questions like "what will Alice say next?" but not questions like "is Alice actually happy with the situation?" It's built up a lot of knowledge about the world, and uses that knowledge to estimate accessible facts well as it can.

We suspect there is a related model M* that uses all of M's machinery in order to explain what's going on in Alice's head. By the definition of inaccessibility, M* is at least somewhat more complex than M. So if we want to learn M*, we need a training objective that penalizes M relative to M*.

But it's really easy to tell M* and M apart! M won't automatically tell us what Alice is thinking, but it also won't automatically make up credible stories that hold together. Both of those require extra complexity. So we can still hope to get M* if we train our model to (i) correctly provides accessible info, (ii) says something consistent and plausible about inaccessible info.

The real problem comes from what I'll call the *instrumental policy.* Let's say we've tried to dream up a loss function L(x, y) to incentivize the model to correctly answer information we can check, and gives at least plausible and consistent answers on things we can't check. By definition, the values L(x, y) are themselves accessible.

Then it's natural to learn a policy like: "on input x, produce the output y for which the loss L(x, y) will be minimal." Let's write BAD for this policy.

Put differently, BAD is the model that performs a simple substitution—when you ask it question Q, it replaces it with the meta question "What answer to question Q will make the person training me happiest?" By definition, this new question is also accessible, so it can be posed to M.

BAD may be much simpler than M*, because all it has to do is describe this substitution, whereas M* needs to build new machinery to explain inaccessible facts.

If M is very weak, BAD may perform much worse than M* because it doesn't understand the loss function very well. But if M understands a lot about the world, and about its training process, then it seems like BAD will get at least as low a loss as M*—it's using the same machinery as M and M*, and picking whatever value leads to the lowest loss. It can choose to report its own best estimates if that's the best strategy, e.g. for questions where we can check. So by the time M and M* are human level, it already seems like BAD should be getting as low a loss as M*.

This gets a little bit more complicated if I do consistency checks or if I have large output spaces, and of course humans aren't *that* good at lying, but it seems like the basic idea is pretty solid: once we have a model of accessible information, we can directly turn it into a model that will pass whatever checks we throw at it.

For concreteness, imagine M is a model of Alice's behavior and we are asking it "is Alice actually happy?" Then M* is the model which translates its model of Alice's psychology into words and explains it to us, while BAD is the model that asks "What answer would sound most believable to the person asking the question?"

The question is: if we start with model M and fine-tune it to answer inaccessible questions, is it easier to learn to articulate what it believes or is it going to learn to answer strategically?

We need to be *very* careful about our checks if we want to ensure that BAD doesn't get a lower loss than M*. For example, it may turn out that sometimes the truth looks a little bit wrong to us. . . . And if we do everything right, then M* and BAD perform equally well, and so we may not have much control over which one we get.

IV. When inaccessible info is a safety problem

Let's get a bit more detailed about the argument in section II. I think that our inability to access inaccessible info would become a safety problem when:

1. We care about inaccessible facts, so we can't just evaluate plans based on their accessible consequences.
2. Inaccessible info is a competitive advantage—agents who are blind to inaccessible facts about the world will get outcompeted.
3. There are *some* agents who are able to use inaccessible facts to acquire influence, e.g. because they are optimizing accessible long-term goals.

1. We care about inaccessible facts

If I only cared about accessible facts, then I might not need to ever access inaccessible facts. For example, if I cared about my life expectancy, and this was accessible, then I could ask my AI "what actions lead to me living the longest?" and execute those.

For better or worse, I think we are likely to care about inaccessible facts.

- Generally we care about what's *actually happening* and not just what appears to be happening. We don't want smiling faces on cameras. And if there's a lot of inaccessible action in the world, then it's reasonably likely for accessible indicators to be systematically manipulated by inaccessible forces.

- We care intrinsically about what happens inside people's heads (and inside computers), not just outward appearances. Over the very long term a *lot* may happen inside computers.
- If we totally give up on measuring how well things are going day-to-day, then we need to be actually optimizing the thing we really care about. But figuring that out may require reflecting a long time, and may be inaccessible to us now. We want a world where we actually reach the correct moral conclusions, not one where we believe we've reached the correct moral conclusions.
- Our real long-term priorities, and our society's long-term future, may also be really weird and hard to reason about even if we were able to know what was good. It just seems really bad to try to evaluate plans only by their very long-term consequences.
- We care about things that are far away in space or time, which I think are likely to be inaccessible.

Overall I'm quite skeptical about the strategy "pick an accessible quantity that captures everything you care about and optimize it." I think we basically need to optimize some kind of value function that tells us how well things are going. That brings us to the next section.

2. Inaccessible info is a competitive advantage

Instead of using AI to directly figure out whether a given action will lead to human flourishing over the coming centuries, we could use AI to help us figure out how to get what we want over the short term—including how to acquire resources and flexible influence, how to keep ourselves safe, and so on.

This doesn't require being able to tell how good a very long-term outcome is, but it does require being able to tell how well things are going. We need to be able to ask the AI "which plan would put us in an *actually good* position next year?"

Unfortunately, I think that if we can only ask about accessible quantities, we are going to end up neglecting a bunch of really important stuff about the situation, and we'll be at a significant competitive disadvantage compared to AIs which are able to take the whole picture into account.

As an intuition pump, imagine a company that is run entirely by A/B tests for metrics that can be easily checked. This company would burn every resource it couldn't measure—its code would become unmaintainable, its other infrastructure would crumble, it would use up goodwill with customers, it would make no research progress, it would become unable to hire, it would get on the wrong side of regulators…

My worry is that inaccessible facts will be similarly critical to running superhuman businesses, and that humans who rely on accessible proxies will get outcompeted just as quickly as the company that isn't able to optimize anything it can't A/B test.

- Even in areas like business that society tries particularly hard to make legible, evaluating how well you are doing depends on e.g. valuing intellectual property and intangible assets, understanding contractual relationships, making predictions about what kinds of knowledge or what relationships will be valuable, and so on.
- . In domains like social engineering, biology, cybersecurity, financial systems, *etc.*, I think inaccessible information becomes even more important.
- If there is a lot of critical inaccessible information, then it's not clear that a simple proxy like "how much money is actually in my bank account" is even accessible. The only thing that I can directly check is "what will I see when I look at my bank account statement?", but that statement could itself be meaningless. We really care about things like who effectively controls that bank account and what would really happen if I tried to spend the money. (And if I largely care about inaccessible facts about the world, then "what would happen if I tried to spend my money?" may itself be inaccessible.)
- I can pay inaccessible costs for an accessible gain—for example leaking critical information, or alienating an important ally, or going into debt, or making short-sighted tradeoffs. Moreover, if there are other actors in the world, they can try to get me to make bad tradeoffs by hiding real costs.

3. Some AIs can plan with inaccessible info

So far this discussion could just be about an *AI missed opportunity*, not an *AI risk*.

Things become problematic when it is possible to build AI systems that do use inaccessible info to pursue ambitious long-term goals that would conflict with human flourishing. If illegible knowledge is important enough, those systems could outcompete humans and divert some (or almost all) of our civilization's resources.

This happens if *any* interesting long-term goal is accessible, i.e. if there's any accessible goal that benefits from accumulating influence.

Why might some long-term goal be accessible?

- Verifiable long-term predictions may accessible, e.g. "total reward over the next 50 years" might be accessible as a generalization of "total reward over the next 50 minutes."
- Even if models don't transfer in a natural way to very long time horizons, they may still transfer in *some* way. Maybe they can't generalize to something as cleanly specified as "total reward," but e.g. they may be

able to predict quantities that happen to be natural in terms of the representation they are using.
- More generally, models will do *something* when we run them way out of distribution. I think the best reason to expect models to not plan intelligently over long time horizons is that they lack the abilities and knowledge. Once they have relevant cognitive machinery, I think it's pretty likely that they'll use the machinery somehow when they go out of distribution, and if you search over a lot of possibilities.

The key asymmetry working against us is that optimizing flourishing appears to require a particular quantity to be accessible, while danger just requires *anything* to be accessible.

There's a longer discussion about whether anyone would want to build such AI systems, or might be likely to do it accidentally. And another discussion about how bad it would actually be. But I think that once we satisfy the 3 properties in this section then we have a real problem, and the remaining question is one of magnitude, with a bunch of factors pulling up and down.

V. Possible responses

At a high level, we'd like to get good answers by training a model to make checkable predictions and then somehow incentivizing it to use the same machinery to also report inaccessible facts, for example by forcing it to report views and enforcing consistency.

I think there are a bunch of problems with that plan, but the strongest one is that it's totally unclear how we ever get the intended policy rather than the instrumental policy.

Here are some possible angles of attack:

- In some intuitive sense there is a discontinuity between saying X because you believe it and saying X because you believe that it's what you need to say to get a high reward. If weak policies are straightforward rather than instrumental, then maybe we can exploit continuity to avoid the instrumental strategy.
  I'm generally pretty scared of these approaches since they feel pretty brittle, and people seem prone to being optimistic even before we can test or understand anything.
  But I'm a bit more hopeful if we consider continuity of behavior rather than continuity in the space of policies, and if we exploit it by using a model at time T to help check the behavior of a model at time T+1.
- If we were able to actually understand something about what the policy was doing, even crudely, it might let us discriminate between instrumental and intended behavior. I don't think we have any concrete proposals for how to understand what the policy is doing well enough to make this

distinction, or how to integrate it into training. But I also don't think we have a clear sense of the obstructions, and I think there are various obvious obstructions to interpretability in general that *don't* apply to this approach. So together with the other points on this list—especially the existence of a discontinuity and the use of previous versions of a model to help—I think it could be part of a solution.

- The "instrumental" strategy—"pick the answer that minimizes my training loss"—may have a shorter description, but it seems to be unnecessarily computationally complex. First it figures out what's true, and then it strategically decides what to say in light of that. It would be a bit cheaper just to actually report what's true, if we set up the training process well enough that honest reporting got you optimal reward.
  It's not clear how to leverage this though, since "use the simplest model" seems like an important and useful prior. We can't just replace it with "use the cheapest model" without sacrificing competitiveness. That said, it may be able to buy us something in combination with the previous approaches. (See discussion here and here.)
- We aren't able to check inaccessible answers, and they aren't output by the simplest model, but maybe we can make being straightforward the easiest strategy in some other sense. For example, if you have two models who can't interact but need to agree on a consistent story, you could imagine the truth being the easiest thing for them to coordinate on. I think that scheme in particular is pretty unlikely to work, but there is a large space of possibilities here, especially when combined with the other points on this list. For example, maybe you could perform consistency checks between models at different points in time.
- Maybe we can work with models M that don't understand enough about the world, or especially about their training process, in order to turn into the instrumental policy BAD. I tend to be pretty skeptical of "security by obscurity"—if we want to make life hard for the instrumental policy not working, I want it to be because we have randomness hidden from it, or we've constructed a smarter agent that it can't simulate. I think it's pretty bad news if we are relying on a smart model not understanding facts about its own situation, and that's the kind of solution that is particularly likely to break down over time.
- We could hope that our models don't build up important inaccessible knowledge, e.g. because it's possible to justify most interesting conclusions with deduction or because we have reasonably good accessible proxies for our value function. I'm pretty skeptical about this over the long term, but I'm not sure exactly how bad it will be how early.
- The argument in this post is pretty informal, and there's a reasonable chance that you can drive a solution through one of the many gaps/loopholes. I like the problem-solving strategy: "write out the proof that there is no solution, and then sift through the proof looking for a fatal hole."

Overall I don't see an obvious way forward on this problem, but there are enough plausible angles of attack that it seems exciting to think about.

VI. How this relates to amplification and debate

Overall I don't think it's very plausible that amplification or debate can be a scalable AI alignment solution on their own, mostly for the kinds of reasons discussed in this post—we will eventually run into some inaccessible knowledge that is never produced by amplification, and so never winds up in your distilled agents.

In the language of my original post on capability amplification, the gap between accessible and inaccessible knowledge corresponds to an obstruction. The current post is part of the long process of zooming in on a concrete obstruction, gradually refining our sense of what it will look like and what our options are for overcoming it.

I think the difficulty with inaccessible knowledge is not specific to amplification—I don't think we have any approach that moves the needle on this problem, at least from a theoretical perspective, so I think it's a plausible candidate for a hard core if we fleshed it out more and made it more precise. (I would describe MIRI's approach to this problem could be described as despair + hope you can find some other way to produce powerful AI.)

I think that iterated amplification *does* address some of the most obvious obstructions to alignment—the possible gap in speed / size / experience / algorithmic sophistication / etc. between us and the agents we train. I think that having amplification mind should make you feel a bit less doomed about inaccessible knowledge, and makes it much easier to see where the real difficulties are likely to lie.

But there's a significant chance that we end up needing ideas that look totally different from amplification/debate, and that those ideas will obsolete most of the particulars of amplification. Right now I think iterated amplification is by far our best concrete alignment strategy to scale up, and I think there are big advantages to starting to scale something up. At the same time, it's really important to push hard on conceptual issues that could tell us ASAP whether amplification/debate are unworkable or require fundamental revisions.

Figure 4:

Inaccessible information was originally published in AI Alignment on Medium, where people are continuing the conversation by highlighting and responding to this story.

"