AI Alignment Unwrapped

1. Universality Unwrapped
2. Infra-Bayesianism Unwrapped
3. Epistemology of HCH
4. Epistemological Framing for AI Alignment Research

Universality Unwrapped

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

Introduction

Informally, a universal system is universal with respect to any computation; and it is a universal system with respect to a given computation if it understands every set of beliefs that can be ascribed to the computation. The intuition is that the system can reverse engineer most or all of the computation, in order to monitor it or imitate it. This in turn has important consequences for questions of alignment and competitiveness. Universality is the property that defines a universal system. And it is the point of this post.

Universality tries to capture a property needed for many alignement schemes. It was proposed by Paul Christiano, the mind behind many approaches and ideas in the prosaic AGI space, and a founding member of the safety team at OpenAI. Rohin Shah dedicated a full Alignment Newsletter to covering all 6 posts on Universality. Rohin and Evan Hubinger, two important researchers in this field, consider Universality as one of the most exciting research idea of the last few years.[1] Yet nobody talks about Universality.

Except for the Alignment Newsletter mentioned above and a response post by Evan, nothing in the Alignment Forum addresses this idea. I've seen no great discussion, no debates, no counter-arguments or criticism. The original post on Medium has no comments, and the crossposted version here only has a handful, mostly asking for clarification. And the other posts in the sequence rely on understanding this first.

The simplest solution to this problem is to tell you to read the original post. Unfortunately, it is as dense as Q in R, brimming with ideas, intuitions, semi-formal explanations and the many meanderings that research takes before arriving on solid ground. That is to say, you'll have to work for it. Not everyone who might benefit from an understanding of Universality has the time, the need or the want for such an upfront investment.

This post endeavors to be the next best thing: an unwrapping of the main post on universality, from the perspective of one who already took the time to mine it for its insights. Because I started at the same point as you -- or even below -- our inferential distance is hopefully smaller than the one you have with Christiano. And armed with this clarification, you should also be able to tackle the next posts in his sequence.

Before digging into Universality itself, I present the perspective from which I'm reading Christiano's original post: that it's really about understanding computations, and that the main contribution lies in posing the right question instead of the partial answer proposed. I then follow by an explanation of the intuitions behind universality, notably what it is (a property about epistemic domination), why it matters for AI alignment (competitiveness and forcing honesty), and examples of ways to be universal for concrete classes of computations. Then, and only then, I detail Christiano's proposal as a definition of Universality: Ascription Universality. Finally I conclude by giving open problems raised by the post, and wrap up with a summary of the takeaway ideas.

Thanks to Paul Christiano, Evan Hubinger, Jérémy Perret and Rohin Shah for feedback.

How to read the Universality post

I first read the original post after watching a Q&A where Rohin praised it as one of the ideas that excited him the most in AI Safety. Although I didn't grasp everything after this read, I thought I had the gist of it: the post talked about this formal property called Ascription Universality, which would ensure that a system with this property would beat other computations at their jobs.

I was wrong.

So that you don't repeat my mistake, let me prime you before explaining the post further: **Christiano's main point is the proposal of an open problem about understanding computations.**

First, the gist of the post lies in the problem, not in the partial solution. This is made harder to see because the problem is not well-defined. It isn't Fermat's Last Theorem, or the relation of P with NP. Instead Universality is what I call an **open theory problem**. It doesn't ask to solve a concrete and well specified problem; instead, it asks us to find a definition, a concept, a theory that captures a list of intuitions. Other examples are Goal-directedness and Abstraction.

So the point of the post is to present the intuitions behind Universality, as well as its value for AI safety. The attempt at a solution shows how one could go about it, points to some problems and makes the discussion more concrete. But it should not be confused with the theme, which is the open theory problem of Universality. As a corollary, it matters more to get the wobbly part about the intuitions than the specific mathematical details of the partial solution. The structure of my explanation reflects this: I present almost everything at the level of Universality itself, before going into the weeds of Ascription Universality at the end.

The second point is that Universality should be seen as "Universal Understanding": understanding how a system or computation works and why and what it will do.

Why Understanding? Because the concept Christiano is aiming at captures the idea of *knowing as much, or more* than a specific computation. Knowledge is power, especially for computations -- but the point is the knowledge. A system is universal for a computation if, for whatever knowledge or beliefs that can be ascribed to this computation in a "reasonable" way, our system already knows about it. In each case, the universal system must know the knowledge encoded in the computation, which implies it can supervise it and outperform it.

In summary, Christiano's post presents and fleshes out an open theory problem about the ability of some system to completely understand anything useful about some computations.

My position is that this is the clearest and most useful way to read Christiano's post. I make it explicit here both to prime you and to let you backtrack any disagreement to this initial bias I'm committing to. With that said, the rest of the post will not discuss this choice any further.

Universality: Intuitions, Value and Examples

Intuitions about Universality

I proposed in the previous section that Universality is an open theory problem. As such, it consists in a set of intuitions for which the unifying formalization is lacking. Let's explore these intuitions.

Imagine that you have an overseer -- a system which looks at computations for signs of trouble. For example a debate about a concrete neural network, or the amplified supervisor of Iterated Amplification. Then a natural requirement is for the overseer to be able to understand everything that the computation does and understands. This would make the overseer universal in a very intuitive way.

What do I mean by understanding a computation? This is another question in need of formalization. What Christiano gives is an intuition and a sort of extensive definition. Understanding a computation means intuitively to understand all beliefs of the computation -- everything that it knows. Examples of such beliefs are:

- The final decision of the computation
- The recognition of a dog’s nose at a layer of a convolutional neural network
- The fact that the computation pretends to act aligned before taking a treacherous turn.

So beliefs in this sense capture all the information inside a computation. This includes both the information that the computation gives us (its output for example) and the information it doesn’t give us (like deceptive intent or any inaccessible information).

Yet what does it mean for information to be hidden inside a computation? Here Christiano doesn’t pretend to extract the correct beliefs of the computation, but instead enlarges his requirement to any reasonable ascription of beliefs to the computation. For any way to ascribe beliefs and knowledge to a specific computation that makes sense and isn’t too strong, this constitutes something that a universal system for this computation must get.

Literary interpretation offers a helpful analogy here. In "The Limits of Interpretation,” Umberto Eco says that any interpretation of a text is valid as long as it survives contact with the text. The interpretative act aims not at finding exactly what the author meant -- usually a hopeless endeavor -- but instead to find interpretations which survive falsification by the text. In the words of Eco himself:

> We can thus accept a sort of Popper-like principle according to which if there are not rules that help to ascertain which interpretations are the "best ones," there is at least a rule for ascertaining which ones are "bad." This rule says that the internal coherence of a text must be taken as the parameter for its interpretations.

A reasonable ascription of beliefs works in the same way: the beliefs should not contradict the actual computation, should explain it and shouldn’t be too strong in a way that is not justified by the computation itself. For such beliefs, any universal system for this computation needs to understand them.

This is not a definition of a reasonable ascription; instead it is a powerful intuition giving us some way to analyse an ascription of beliefs to a computation. As an example, reasonable doesn't mean that we would have thought of it, or even that it's one way we would solve the problem addressed by the computation. A reasonable ascription is not a common-sense ascription, but an ascription that makes sense. In addition with this intuition, we have examples of reasonable ascriptions (the intuitional stance, neural circuits, ...) and unreasonable ones (ascribing all consequences of known facts as beliefs -- logical omniscience).

To summarize the previous discussion: **a universal system with respect to a given computation is a system that understands every set of beliefs that can be ascribed to the computation by any “reasonable” ascription approach.**

A natural requirement would be to ask for a universal system with respect to any computation. But this is obviously impossible: one can always create more complex systems with more complex goals and beliefs, such that any fixed system is just too basic to understand them. The alternative proposed by Christiano is a parameterized notion of universality. In essence, the algorithm used by the universal system for a computation C will depend explicitly on C.

In my first draft, I assumed that the parametrization meant that the algorithm would change for every C or class of C. But Christiano corrected me in his feedback, clarifying that the parametrization plays a role in the resources used by the universal system.

With this out of the way, there are two natural ways exists to do parameterization:

- An abstract parametrization (which Christiano calls idealized), where the parameter is a whole class of computations. For example, a universal system whose running time grows with the number of layers of a neural network.
- A concrete parametrization (which Christiano calls practical), where the parameter is the specific computation, possibly including how it was created, like a training process. For example, a universal system taking as a parameter the training process and the weights of the specific neural network.

Ideally, we want the most abstract possible version of universality, as it would work for entire classes of computations at a time; yet understanding a computation seems far easier with access to the computation itself, even more if the training process is included. This shows a fundamental tradeoff between the generalizability of a universal system and its ability to tailor itself to a specific computation. In some sense, Universality asks the question of how much we can understand with a specific amount of resources (be it time or information about the specific computation), and what is the minimal amount of resources needed to understand the computation fully.

To summarize the complete intuition: **a universal system is parametrized by the computation it needs to understand, either in a loose way for the idealized version, or in a tight way for the practical version. Understanding a computation means understanding any beliefs that can be reasonably ascribed to the computation -- any beliefs that explain and justify it while not contradicting the computation.**

One point this fails to address is what understanding means -- that depends on the specific solution given to Universality. So I'll get back to this question when exploring Ascription Universality, the partial solution proposed by Christiano to the open problem of Universality. But we can do for now by imagining that understanding means that the universal system can explain to us the set of beliefs.

Why care about Universality?

Suppose that we solve the open theory problem of universality: we find a formal definition of a universal system. What does it buy us?

To find out, let's turn to interpretability. Modern research in ML includes many techniques to look through a trained program and extract information from it. Formal methods are now also able to check properties of trained systems, just as they do for more traditional programs. But what are we trying to check, to find out? It's fundamental: interpretability and formal methods without a property to check are just a formal game without any useful consequence.

Solving Universality gives a property to check for two guarantees in an overseer: its competitiveness and its ability to make computation honest.

Most alignment schemes add constraints on top of modern methods of optimization and/or ML. It's thus sensible to ask about their cost. The word used in the discussion tends to be competitiveness. Now, there are numerous versions of competitiveness, with classification that disagree with each other, but the only one we care about here is performance competitiveness. Can the approach actually be as good as the non-aligned computations? Universality shifts this question, by asking instead: if we had a (potentially unsafe) computation that is competitive for a given task, can we reverse-engineer why the computation works? If we can, then implementing the same strategy means arriving at the same performance. Thus universality would be strong evidence in favor of competitiveness.

Dealing with honesty, the second use of Universality, shows the value of asking about knowledge and beliefs instead of just about beating the computation. A universal system, by knowing the beliefs and knowledge that can be ascribed to a computation, would be able to understand everything the model knows, not just what it displays. The overseer would thus spot deception, see problems of inner alignment, and in general extract all the inaccessible information hidden in the computation.

Hence Universality captures, in one idea, both the ability to outperform computations and to "see through" them.

Example of Universality

Thinking about Universality by itself, without confronting it to specific computations, holds the risk of missing a fundamental intuition. So Christiano gives examples of powerful and interesting classes of computations, with his best guesses on how to be universal for them. Note that every example here assumes an access to the computation itself, and usually even to the training data. This means they fall under the concrete (or practical) parametrization of Universality.

Here is Christiano describing the pattern of his examples.

The first three examples will consider programs C with the following form:

- Iterate over programs to find an f that successfully predicts the answers to some dataset of historical questions for which we know the answer.
- Use that program to predict the answers to future questions.

The difference lies in the inductive bias of the search -- the kind of f this program is looking for. An important assumption here is that f will be competent at its task. Then Christiano considers three inductive biases:

- **Deduction** f is a deduction process which receives a question and a context, deduces information about the answer from the context, and then answers in a way coherent with its deductions.
- **Modeling** f converts a question and a context to a physical situation, which it models and runs until the state corresponding to the answer is reached.

- **Alien Reasoning** f uses a complex reasoning procedure that we don't really use naturally. In the post, it builds a matrix with specific properties in order to predict the stock exchange.

Christiano uses HCH (Humans Consulting HCH, the limiting behavior of IDA) as his placeholder universal systems, but only requires that it works as a strong bayesian. If it does, then he proposes a similar approach to solve all three case:

- First, the strong bayesian notices that f gives good answers (using for example the training dataset and observations of performance in production)
- Depending on the procedure, it then infers that using it gives good answers, which by bayesian reasoning means that the specific assumptions of f (the axioms for a deduction process, the physical model for a modeling problem) are likely to be true.
- Then the strong bayesian leverages the process used by f to derive the same answers.

For the deduction process, it might be straightforward as deductions could already live at the same level of abstraction than the questions and answers. But for both modeling and alien reasoning, an additional step is probably necessary: finding a bridge hypothesis between the internal state of f and the ascribed beliefs in terms of what it tries to predict. For the modeling, it might mean interpreting the low-level physical system at the level of humans. For this specific alien reasoning, it means inferring the link between the matrix and the quantity to predict. The universal system might be able to simply search over such bridging hypotheses, and infer from how much it explains whether it might hold or not. Since every coherent bridging hypothesis that explains the predictions of f corresponds to a reasonable ascription of beliefs, such a system would dominate all of them.

Christiano posits another example in his post. It's the most interesting one in my opinion: a computation C which contains a process f, such that f designs a new reasoning process g to solve the problem at hand. Such a computation acts similarly to a programmer thinking of some neat solution, testing it a little bit, and then shipping it.

Because of the lack of inference, a strong bayesian cannot just give enough credence to the process of generating the reasoning process on statistical grounds. Instead, the system trusts f because it works (that part is still statistical); then because f uses g, f has has beliefs on g. Which means that assuming the system is already universal for the rest of f, it should be able to understand the beliefs of f about g.

The Proposal

Now that we explored the web of intuitions underlying Universality, we can look into details into Christiano's partial solution: Ascription Universality. Remember that the point of the post is to ask the open theory problem, not to defend Ascription Universality as the complete answer.

A Subjective Perspective

When giving the intuitions behind Universality, I didn't pin down what understanding the computation's beliefs might mean. This leaves multiple choices when writing an actual definition. The two questions to answer are:

- How to represent the beliefs ascribed to the computation?
- What does it mean for the universal system to understand these beliefs?

Christiano answers the first question by defining beliefs in terms of probabilistic predictions. In the definition of Ascription Universality, They're represented by expectation functions, which give an estimate (the belief) for every random quantity. As for the second question, an intuitive answer would be to make the universal system make more accurate predictions than the computation. That is to say, to have the beliefs of the universal system be more accurate for all variables than those ascribed to the computations.

How do we learn the beliefs of our universal system? Fortunately, Christiano writes in a context where potentially universal systems are question-answerers using a fixed language. We can thus just ask it in English what are its beliefs about something.

So understanding the beliefs of the computation means that the beliefs of the system are more accurate than those of the computations for every variable. There's only one problem: it's impossible! Even a superior system, far more smarter or knowledgeable, is not omniscient. It might make mistakes, and the computation might make lucky guesses.

That is to say, the naive approach to comparing beliefs doesn’t work here. So instead, Christiano proposes a subjective approach: **fixing an epistemic perspective (a set of beliefs), from which to judge the accuracy of the system's beliefs compared to the computation's beliefs.**

Thus a universal system understands the beliefs ascribed to a computation relative to a given epistemic perspective, if that perspective believes the predictions of the universal system to be more accurate than those of the computation.

Of course, a subjective definition raises many questions: isn't it too weak? What guarantees does it provide? I'll get back to these interrogations after the formal definition. But a little teaser: subjective domination (beliefs being more accurate from a fixed subjective perspective) ends up being a stronger condition than it seems at first glance.

The Formal Definition

As explained above, beliefs are represented as probabilistic predictions. Christiano considers beliefs as expectation functions, which gives the expected value for any variable in a possibly infinite set. Another way to see these expectations is as infinite tables of expected values (or infinite vectors). Note that this table includes not only pure expectations like $E[X]$, but also every conditional expectation like $E[X|Y]$.

Now, let's formalize the idea of more accurate beliefs. We have two sets of beliefs represented by the expectation functions $E_{overseer}$ and $E_{computation}$. Because of the subjective bent of this definition, we also have an expectation E_{human} that plays the role of epistemic perspective.

(I changed the names of the expectations in the quote to stay coherent.)

> Then we say that $E_{overseer}$ **dominates** $E_{computation}$ (w.r.t. E_{human}) if, for every bounded quantity X and for every “nice” property Φ:
>
> $$E_{human}[X \mid \Phi(E_{overseer}, E_{computation})] = E_{human}[E_{overseer}[X] \mid \Phi(E_{overseer}, E_{computation})]$$
>
> (By “nice” I mean something like: simple to define and open in the product topology, viewing $E_{overseer}$ and$E_{computation}$ as infinite tables of numbers.)

First, an intuition: $E_{overseer}$ dominates $E_{computation}$ if whatever reasonable information is given to E_{human} about $E_{overseer}$ and $E_{computation}$, E_{human} still trusts the predictions of $E_{overseer}$ more than those of $E_{computation}$.

That's a lot to unwrap.

Boundedness of X This point doesn't matter too much. It's probably just a way to ensure that $E[X]$ exists, without using complex conditions on convergence.

Nice Property Φ What is Φ, to start? Intuitively, it represents the information that E_{human} has on both $E_{overseer}$ and $E_{computation}$. Or said differently, it represents what one could tell E_{human} to make it change one way or another. Its domain is the product of the expectation function space with itself. For its codomain, Christiano doesn't specify it. We can take $\{0, 1\}$ with the discrete topology -- as we'll see, any discrete space will do. In that way, Φ is a function of both beliefs, that returns some information (maybe a bit, maybe more) about the two beliefs.

Christiano also writes that Φ must be "nice". I unwrap the formal conditions in the appendix -- spoilers: it takes some work -- but the intuition couldn't be simpler: **a nice property is one which depends only on the expectation of a finite number of variables**. In that sense, Φ truly represents information that could be given to E_{human} to sway it.

Once we have the domination of beliefs clarified, the rest of the solution is simple: a parameterized system A[.] is **ascription universal** from the perspective of E_{human} if, for all computations C, the beliefs of A[C] dominates, according to E_{human}, any reasonable ascription of beliefs to C.

Why is this only a partial solution? First, it is not completely formal: it lacks a definition of reasonable ascription procedures. Even the standard ascription of beliefs to A[C] is hard to manipulate from a formal perspective, despite being more defined. And second, there are subtleties with the subjective perspective that needs to be resolved.

Discussing the Definition

There are three points that need to be discussed about Ascription Universality: subjectivity and opacity.

Subjectivity I pointed out multiple times that the definition of Ascription Universality proposed by Christiano is subjective. At first, it was bugging me. After all, even if we prove universality according to our perspective, what guarantees does that actually offer? Maybe we are stupid. Or just not smart enough to get it. That would mean it doesn't buy us anything.

Then Christiano writes this:

> In applications of this property, the subjective condition is what we need in order for us to believe that A[C] will cope with the challenges posed by C. For example, suppose that C formulates a plan to “trick” A[C]. Then the subjective universality condition implies that we don’t expect C to succeed.
>
> This isn’t as good as actually knowing that C won’t succeed. But I think it should be good enough for us — the reason we are thinking about AI safety is because we are concerned that something bad will happen. If we find a technique

> that defuses this argument, then we’ve addressed the motivating problem. It may still be the case that bad things happen (and we should still search for additional reasons that bad things might happen), but we don’t particularly expect them to.

I don't know about you, but I first read this as saying "as long as humans can't see the risk, **even if they could by being smarter or more thorough**, then we don't care about the risk". A statement with which I disagree wholeheartedly.

But then Evan Hubinger reminded me that here, Ascription Universality means that humans can't see the risk **whatever finite information is given to them about the beliefs of A[C] and C**. That is far stronger. It means that whatever research we do, we wouldn’t find anything convincing us of the existence of the risk.. I'm still not sure it's enough (it depends on the epistemic perspective), but now it doesn't seem trivially wrong.

Actually, it asks a couple of fascinating questions:

- Given a human being, for which deceptive C is there a Φ allowing the human to find out the deception?
- Given an epistemic perspective E, for which deceptive C is there a Φ allowing E to find out?

Opacity Another criticism that Christiano attempts to nip in the bud is that Universality doesn't require an understanding of the computation. Talking about an example computation which search programs to classify images, he write:

In this case, we can ascribe beliefs to C about the contents of the new image. And because those beliefs are coming from a simple program that works empirically, I expect them to be accurate (in some respects).

For example, a simple classifier C may “believe” that the new image contains a particular curve that typically appears in images labeled “dog;” or a really sophisticated classifier may perform complex deductions about the contents of the scene, starting from premises that were empirically validated on the training set.

So basically, there must be reasons for which the heuristics used by C works. These reasons then translate into beliefs which a universal system must understand, and thus it must understand how the heuristics work.

I'm sympathetic with this intuition. My only caveat is that it relies on a conjecture: that every good heuristic admits a simple enough explanation. I believe it to be true, but I still want to point out the reliance of this argument on it.

Open Problems

Last but not least, I promised a list of open problems. Some papers in theoretical computer science (like those of Scott Aaronson) end with a list of the open problems that feel exciting to the authors. I really like that, because it gives me a jumping point to go further and try to push this research direction. So this list extracts all the open problems I could find in this post. I also separated them into open theory problems and open concrete problems, where the latter are what would usually be called open problems about Ascription Universality.

Open Theory Problems

- Is there an objective definition for Universality, which captures the intuitions in this post?
- How do we define the set of reasonable ascriptions of beliefs to a computation C?
- Is there a simpler, easier to use definition of Universality leveraging some constraint on the structure of C?
- When is an epistemic perspective for Universality sufficient for AI alignment?
- What is the evidence that we can build universal systems?

Open Concrete Problems (for Ascription Universality)

- If an idealized system is universal, can we implement a practical version that stays universal?
- Given a human being, for which deceptive C is there a Φ allowing the human to find out the deception?
- Given an epistemic perspective E, for which deceptive C is there a Φ allowing E to find out?
- For what formally specified C and A[C] does Ascription Universality hold?
- What other conditions might be equivalent to Ascription Universality in different circumstances?

Conclusion

Universality is the sort of problem that guides theory research. It posits that behind our intuitions for beating a computation and forcing it to be honest, there’s a common thread which can be abstracted away. Armed with this property, we could use testing, formal verification, and interpretability to extract guarantees about alignment schemes.

Christiano’s original post (and the concurrent ones) gave this problem to the field. What we need now is people looking into it, toying with it, and unearthing parts of answers.

Appendix

Remember that Φ must be "nice" in the definition of Ascription Universality. I wrote above that a nice property is one which depends only on the expectation of a finite number of variables.

In the definition, Christiano asks for Φ to be an open function. Yet I think that instead, he want Φ to be continuous, as written a bit later:

> (continuity in the product topology is the minimum plausible condition to avoid a self-referential paradox)

A fundamental topological property of continuous functions is that the preimages (the sets of points whose image by the function is in the given set) of open sets are open. Back in our definition, notice that the domain of Φ is a discrete space, such that $\{0\}$ and \{1\} are both open. Continuity of Φ then entails that the preimages of $\{0\}$ and $\{1\}$ by Φ are open.

That is to say, the sets of expectations for which Φ returns a fixed variable are open sets. This put a constraint on them, which explains the intuition behind a nice property.

The last piece of the puzzle is the product topology. Or to be exact, two meanings of the term "product topology": the induced topology on a product space by the topology of the building blocks of the product; and the standard topology on function spaces.

Because the domain of Φ is a product of two function spaces, the obvious topology to apply to it is the product topology: the topology whose open sets are the products of open sets in the two topologies.[2] But what are those topologies of the function spaces?

Now, there are many possible topologies on function spaces. But the one that makes sense here is called... the product topology. How practical. The definition of the product topology for functions from A to B relies on a subbasis to define all its open sets. A subbasis build all the open set by taking all finite intersections among itself, and then taking all unions among these finite intersections. There's thus a real sense in which a subbasis spans a topology.

The subbasis of the product topology (for functions from A to B) has an element for every element a of A and every open set U of B: $S(a, U) = \{f \in A \to B | f(a) \in U\}$. That is, the set of functions whose value for a is contained in U. Notably, this definition only constrains f at one point, even if A is infinite.

Now, recall that to get the set of all open sets (the topology) from a subbasis, one needs to take all finite intersections of elements of the subbasis. Which, given the form of the subbasis, means that these intersections only constrain the functions at a finite number of values. And we get back our initial condition.[3]

So in summary, Φ **must be continuous so that the sets that are sent to** 0 **and** 1 **by it are open, because open in the corresponding topology means only constraining the functions at a finite number of values**.

[1] ^

Evan in a personal discussion, and Rohin as an answer to a question in a Q&A session for the AI Safety Camp Toronto

[2] ^

Technically, an open set in the product topology is a product of open sets such that only finitely many of these open sets are not equal to their whole space. But for a product of two spaces, this doesn’t matter

[3] ^

Because an infinite union of open sets is open, some open sets actually talk about all the values, but they do it in a slightly different way than constraining them all together. You can represent each open set as a conjunction of finitely many constraints. Then the problematic open sets would be infinite disjunctions of these conjunctions. They don't require an infinite number of constraints to hold at the same time, but they might force us to check an infinite number of clauses to see if the function is in the set.

Infra-Bayesianism Unwrapped

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

Introduction

Infra-Bayesianism is a recent theoretical framework in AI Alignment, coming from Vanessa Kosoy and Diffractor (Alexander Appel). It provides the groundwork for a learning theory of RL in the non-realizable case (when the hypothesis is not included in the hypothesis space), which ensures that updates don't throw away useful information, and which also satisfies important decision-theoretic properties playing a role in Newcomb-like problems.

Unfortunately, this sequence of posts is really dense, and uses a lot of advanced maths in a very "textbook" approach; it's thus hard to understand fully. This comes not from the lack of intuition (Diffractor and Vanessa are both very good at providing intuitions for every idea), but from the sheer complexity of the theory, as well as implicit (or quickly mentioned) links with previous research.

Thus my goal in this post is to give enough details for connecting the intuitions provided and the actual results, as well as the place of Infra-Bayesianism within the literature. I will not explain every proof and every result (if only because I'm not sure of my understanding of all of them). But I hope by the end of this post, you have a clearer map of Infra-Bayesianism, one good enough to dig into the posts themselves.

This post is splitted into three section

- Section 1 explores the context of Infra-Bayesianism, the problem it attempts to solve, and some of the relevant literature. You can see it as unwrapping this section in Introduction to The Infra-Bayesianism Sequence.
- Section 2 gives a bird's-eye view of Infra-Bayesianism; a map to navigate through the sequence.
- Section 3 follows one path through this map: the one focused on decision-theoretic properties for Newcomb-like problems.

Reading advice: The reader is king/queen, but I still believe that there are two main ways to read this post to take something out of it:

- Read it quickly, in one go to get a general idea of Infra-Bayesianism and a very high-level map.
- Read it in detail, for a time between 2 hours and a whole afternoon, to get every detail explained here. If you do so, I really believe that you'll have a quite detailed map of Infra-Bayesianism, enough to explore the sequence by yourself without getting lost in the mathematical jungle.

Thanks to Vanessa and Diffractor for going above and beyond in answering all my questions and providing feedback on this post. Thanks to Jérémy Perret for feedback on this post.

Section 1: Why is Infra-Bayesianism Important?

Cruxes

Before going into the problem tackled by Infra-Bayesianism, I want to give some context in which to judge the value of this research.

AI Alignment is not yet a unified field; among other things, this means that a lot of researchers disagree on what one should work on, what constitutes a good solution, what is useful and what isn't. So the first thing I look for when encountering a new piece of AI Alignment research is its set of underlying assumptions. In rationalist parlance, we would say the cruxes.

Infra-Bayesianism, just like most of Vanessa's research, relies on three main cruxes made explicit in her research agenda:

- **(AI Alignment requires more than experimental guarantees)** I would say this one is almost a consensus among AI Alignment researchers. The fact that so many things can go wrong, at scales where we'll be unable to act, or even notice the issue, means that just running experiments and checking that everything is ok isn't enough. Moreover, in some cases if the experiment fails, it's game over. So we want explanations and models for why the AI we build will indeed be aligned.
- **(Such guarantees must come from a mathematical theory of AI Alignment)** This one, on the other hand, is a major source of disagreement. Most researchers agree that having mathematical results for the kind of guarantees we care about would be great -- some are simply pessimistic that such results exist. See for example Rohin's position in this discussion.
- **(The best candidate for such a theory is a theory of RL)** Lastly, this crux is a bit more difficult to put into context, because it relies on the previous, controversial crux. That being said, many of the people unconvinced by the previous crux seem to focus their research effort into prosaic AGI, that is on DeepRL. So working on RL appears rather accepted too.

In summary, Infra-Bayesianism makes sense as a component of a theory of RL, with the goal of proving formal guarantees on alignment and safety. Even if you disagree with some of these cruxes, I feel that being aware of them will help you understand these posts better.

Non-realizability: the heart of the matter

The main idea motivating Infra-Bayesianism is the issue of non-realizability. Realizability is a common assumption on learning tasks, where the thing we are trying to learn (the function to approximate for example) is part of the hypothesis space considered. Recall that because of fundamental results like the no-free-lunch theorems, learning algorithms cannot consider all possible hypotheses equally -- they must have inductive biases which reduce the hypothesis space and order its elements. Thus even in standard ML, realizability is a pretty strong assumption.

And when you go from learning a known function (like XOR) to learning a complex feature of the real world, then another problem emerges, related to embedded agency: the learning agent is embedded into the world it wants to model, and so is smaller in an intuitive sense. Thus assuming that the hypothesis space considered by the learning algorithm (which is in some sense represented inside the algorithm) contains the function learned (the real world) becomes really improbable, at least from a computational complexity perspective.

One important detail that I missed at first when thinking about non-realizability is that the issue comes from assuming that the true hypothesis is one of the efficiently computable hypotheses which form your hypothesis space. So we’re still in the non-realizable setting when you might know the true hypothesis, but it’s either uncomputable or prohibitively expensive.

Going back to Infra-Bayesianism, non-realizability is a necessity for any practical mathematical theory of RL. But as explained in the first post of the sequence, there is not that many results on learning theory for RL:

For offline and online learning there are classical results in the non-realizable setting, in particular VC theory naturally extends to the non-realizable setting. However, for reinforcement learning there are few analogous results. Even for passive Bayesian inference, the best non-realizable result found in our literature search is Shalizi's which relies on ergodicity assumptions about the true environment. Since reinforcement learning is the relevant setting for AGI and alignment theory, this poses a problem.

If you’re like me, you get the previous paragraph, with the possible exception of the part about “ergodicity assumptions”. Such assumptions, roughly speaking, mean that the distribution of the stochastic process (here the real world) eventually stabilizes to a fixed distribution. Which will probably happen, around the heat-death of the universe. So it’s still a very oversimplified assumption, that Infra-Bayesiansm removes.

Now, the AI Alignment literature contains a well-known example of a non-realizable approach: Logical Induction. The quick summary is that Logical Induction deals with predicting logical consequences of known facts that are not yet accessible due to computational limits, in ways that ensure mistakes cannot be exploited for an infinite amount of “money” (in a market setting where predictions decide the “prices”). Logical inductors (algorithms solving Logical Induction) deals with a non-realizable setting because the guarantee they provide (non-exploitation) doesn’t depend on the “true” probability distribution. Equivalently, logical inductors attempt to approximate a probability distribution over logical sentences that is uncomputable, and that has no computable approximation in full.

Building on Logical Induction (and a parallel line of research, which includes the idea of Defensive Forecasting), a previous paper by Vanessa titled Forecasting Using Incomplete Models, extended these ideas to more general, abstract and continuous settings (instead of just logic). The paper still deals with non-realizability, despite having guarantees that depend on the true hypothesis. This is because the guarantees have premises about whether the true hypothesis is inside an efficiently computable set of hypotheses (a convex set), instead of requiring that the true hypothesis is itself efficiently computable. So instead of having a handful of hypotheses we can compute and saying “it’s one of them”, Forecasting Using Incomplete Models uses efficiently computable properties of hypotheses, and say that if the true hypothesis satisfies one of these properties, then an efficiently computable hypothesis with the same guarantees will be learned.

This idea of sets of probability distributions also appears in previous takes on imprecise probabilities, notably in Walley’s Statistical Reasoning with Imprecise Probabilities and Peng’s Nonlinear Expectations and Stochastic Calculus under Uncertainty. That being said, Vanessa and Diffractor heard about these only after finishing most of the research on Infra-Bayesianism. These previous works on imprecise probabilities also don’t deal with the decision theory aspects of Infra-Bayesianism.

Lastly, all the ideas presented for prediction above, from logical induction to imprecise probabilities, provide guarantees about the precision of prediction. But for a theory of RL, what we want are guarantees about expected utility. This leads directly to Infra-Bayesianism.

Section 2: What is Infra-Bayesianism, and What can it do?

Bird's-eye View of Infra-Bayesianism

The main object of Infra-Bayesianism is the infradistribution (Definition 7 in Basic Inframeasure Theory): a set of "pimped up" probability distributions called sa-measures. These sa-measures capture information like the weight of the corresponding distribution in the infradistribution and the off-history utility, which prove crucial for decision theoretic reasoning further down the line (in Belief Functions and Decision Theory). Infradistributions themselves satisfy many conditions (recapped here in Basic Inframeasure Theory), which serves to ensure it's the kind of computable property of environments/distributions that we want for our incomplete models.

Basic Inframeasure Theory, the first technical post in the sequence, defines everything mentioned previously from the ground up. It also brushes up on the measure theory and functional analysis used in the results, as well as show more advanced results like a notion of update (Definition 11) that takes into account what each sa-measure predicted, the corresponding Bayes Theorem for infradistributions (Theorem 6), a duality result which allow manipulation of infradistributions as concave, monotone, and uniformly continuous functionals (Theorem 4), and a lot of others useful theoretical constructions and properties (see for example the section Additional Constructions).

The next post, Belief Functions and Decision Theory, focuses on using Infra-Bayesianism in a decision theoretic and learning theoretic setting. At least the decision theoretic part is the subject of Section 3 in the present post, but before that, we need to go into more details about some basic parts of inframeasure theory.

(Between the first draft of this post and the final version, Vanessa and Diffractor published a new post called Less Basic Inframeasure Theory. Its focus on advanced results means I won't discuss it further in this post)

Maxmin Expected Utility: Knightian Uncertainty and Murphy

Recall that we want to build a theory of RL. This takes the form of guarantees on the expected utility. There's only one problem: we don't have a distribution over environments on which to take the expectation!

As defined above, an infradistribution is a **set** of probability distributions (technically sa-measures, but that's not important here). We thus find ourselves in the setting of Knightian uncertainty: we only know the possible "worlds", not their respective probability. This fits with the fact that in the real world, we don't have access to clean probabilities between the different environments we consider.

As theoretical computer scientists, Vanessa and Diffractor are fundamentally pessimistic: they want worst-case guarantees. Within a probabilistic setting, even our crowd of paranoid theoretical computer scientists will get behind a guarantee with a good enough probability. But recall that we have Knightian uncertainty! So we don't have a quantitative measure of our uncertainty.

Therefore, the only way to have a meaningful guarantee is to assume an adversarial setting: Murphy, as he's named in the sequence, chooses the worst environment possible for us. And we want a policy that maximizes the expected utility within the worst possible environment. That is, we take the maxmin expected utility over all environments considered.

To summarize, we want to derive guarantees about the maxmin expected utility of the policy learned.

From Probability Distributions to Sa-Measures

So we want guarantees on maxmin expected utility for our given infradistributions. The last detail that's missing concerns the elements of infradistributions: sa-measures. What are they? Why do we need them?

The answer to both questions comes from considering updates. Intuitively, we want to use an infradistribution just like a prior over environments. Following the analogy, we might wonder how to update after an action is taken and a new observation comes in. For a prior, you do a simple bayesian update of the distribution. But what do you do for an infradistribution?

Since it is basically a set of distributions, the obvious idea is to update every distribution (every environment in the set) independently. This has two big problems: loss of information and dynamic inconsistency

Relative probabilities of different environments

In a normal Bayesian update, if an environment predicted the current observation with a higher probability than another environment, then you would update your distribution in favor of the former environment. But our naive updates for infradistributions fails on this count: both environments would be updated by themselves, and then put in a set. Infra-Bayesianism's solution for that is to consider environments as scaled distributions instead. The scaling factor plays the role of the probability in a distribution, but without some of the more stringent constraints.

Now, these scaled measures don't have a name, because they're not the final form of environments in Infra-Bayesianism.

Dynamic Consistency

Even with scaled measures, there is still an issue: dynamic inconsistency. Put simply, dynamic inconsistency is when the action made after some history is not the one that would have been decided by the optimal policy from the start.

For those of you that know a lot of decision theory, this is related to the idea of commitment, and how they can ensure good decision theoretic properties.

For others, like me, the main hurdle for understanding dynamic consistency is to see how deciding the best action at each step could be suboptimal, if you can be predicted well enough. And the example that drives that home for me is Parfit's hitchhiker.

> You're stranded in the desert, and a car stops near you. The driver can get you to the next city, as long as you promise to give him a reward when you reach civilization. Also very important, the driver is pretty good at reading other human beings.

Now, if you're the kind of person that makes the optimal decision at each step, you're the kind of person that would promise to give a reward, and then not give it when you reach your destination. But the driver can see that, and thus leaves you in the desert. In that case, it would have been optimal to commit to give the reward and not defect at your destination.

Another scenario, slightly less obvious, is a setting where Murphy can choose between two different environments, such that the maxmin expected utility of choosing the optimal choice at each step is lower than for another policy. Vanessa and Diffractor give such an example in the section Motivating sa-measures of Introduction To The Infra-Bayesianism Sequence.

The trick is that you need to keep in mind what expected utility you would have if you were not in the history you're seeing. That's because at each step, you want to take the action that maximizes the minimal expected utility over the whole environment, not just the environment starting where you are.

Vanessa and Diffractor call this the "off-history" utility, which they combine with the scaled measure to get an a-measure (Definition 3 in Basic Inframeasure Theory). There's a last step, that lets the measure be negative as long as the off-history utility term is bigger than the absolute value of any negative measure: this is an sa-measure (Definition 2 in Basic Inframeasure Theory). But that's mostly relevant for the math, less for the intuitions.

So to get dynamic consistency, one needs to replace distributions in the sets with a-measures or sa-measures, and then maintain the right information appropriately. This is why the definition of infradistributions uses them.

Interestingly, doing so is coherent with Updateless Decision Theory, the main proposal for a decision theory that deals with Newcomb-like or Parfit's hitchhiker types of problems. Note that we didn't build any of the concepts in order to get back UDT. It's simply a consequence of wanting to maxmin expected utility in this context.

UDT also helps with understanding the points of updates despite dynamic consistency: instead of asking for a commitment at the beginning of time for anything that might happen, dynamically consistent updates allows decisions to be computed online while still being coherent with the ideal precommitted decision. (It doesn't solve the problem of computing the utility off-history, though)

Section 3: One Path Through Infra-Bayesianism, Newcomb-like Problems and Decision Theory

Lastly, I want to focus on one of the many paths through Infra-Bayesianism. Why this one? Because I feel it is the most concrete I could find, and it points towards non obvious links (for me at least) about decision theory.

This path starts in the third post of the sequence, Belief Function and Decision Theory

Beliefs Functions and their Unexpected Consequences

Beliefs functions (Definition 11 in Belief Function and Decision Theory) are functions which take as input a partial policy (according to Definition 4), and return a set of a-measures (according to the definitions in Basic Inframeasure Theory mentioned above) on the outcome set of this partial policy (according to Definition 8).

We have already seen a-measures in the previous sections: they are built from a scaled distribution (here over outcomes) and a scalar term that tracks the off-history utility (to maintain dynamical consistency). For the rest of the new terms, here are the simple explanations.

- An o-history h (Definition 3) is a sequence of alternating observations and actions, that ends with an observation. These are the input to policies, which then return the next action to take.
- A partial policy π_{pa} (Definition 4) is a partial function from o-histories to actions, such that π_{pa} is defined coherently with the prefixes of histories on which it is defined: if there is an o-history h such that $\pi_{pa}(h)$ is well defined, then for every prefix of h of the form $h'a$, we have $\pi_{pa}(h') = a$. Basically, a partial policy is defined on increasing o-histories, until it’s not defined anymore.
- The outcome set $F(\pi_{pa})$ (Definition 8), is the set of o-histories that are not in the domain of π_{pa}, but which have all of their prefixes in it, and the output of π_{pa} are the coherent actions for these prefixes. These are the o-histories where π_{pa} stops (or its infinite histories)

To summarize, a belief function takes a policy, which gives new actions from histories ending in observations, and returns a property on distributions over the final histories of this policy. This generalizes a function that takes a policy and returns a distribution over histories.

Now, a confusing part of the Belief Function and Decision Theory post is that it doesn’t explicitly tell you that this set of a-measures over outcomes actually forms an infradistribution, which is the main mathematical object of Infra-Bayesianism. And to be exact, the outputs of a belief function are guaranteed to be infradistributions only if the belief function satisfies the condition listed here. Some of these conditions follow directly from the corresponding conditions for infradistributions; others depend on the Nirvana trick, that we will delve into later; still others are not that important for understanding the gist of Infra-Bayesianism.

So at this point in the post, we can go back to Basic Inframeasure Theory and look at the formal definition of infradistributions. Indeed, such a definition is fundamental for using belief functions as analogous to environments (functions sending policies to a distribution over histories).

Easier Infradistributions: the Finite Case

The general case presented in Basic Inframeasure Theory considers measures, a-measures and sa-measures defined over potentially infinite sets (the outcome set might be infinite, for example if the policy is defined for every o-history). This requires assumptions on the structure of the set (compactness for example), and forces the use of complex properties of the space of measures (being a Banach space among other things), which ultimately warrants the use of functional analysis, the extension of linear algebra to infinite dimensional spaces.

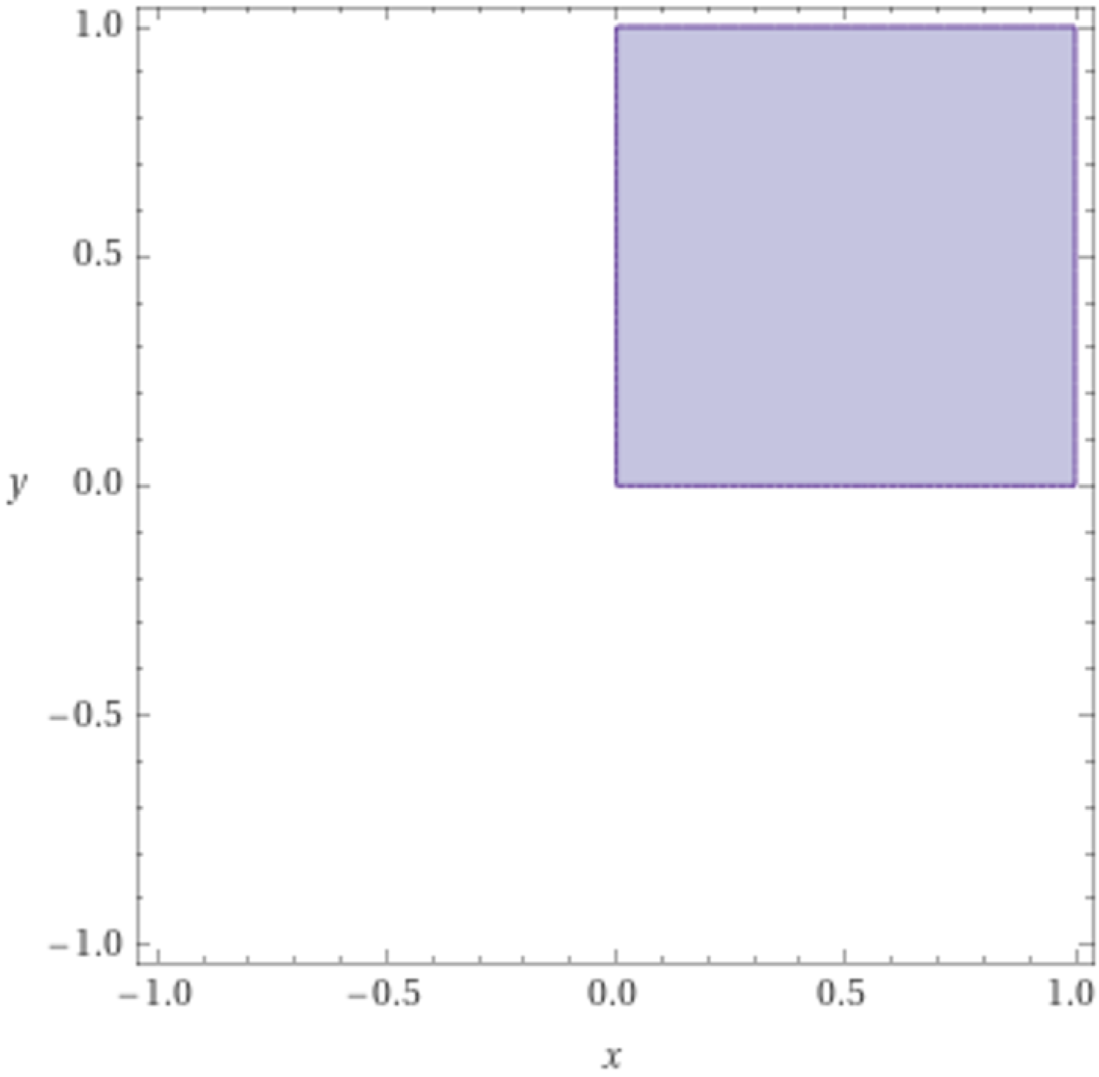
Personally, I’m not well read enough in measure theory and functional analysis to follow everything without going back and forth between twenty Wikipedia pages, and even then I had trouble keeping with the high level abstractions.

Fortunately, there is a way to simplify tremendously the objects with which we work: assume the finiteness of the set on which measures are defined. This can be done naturally in the case of outcome sets, by considering $X_n = F_n(\pi_{pa})$, the set of outcomes of length ≤ n.

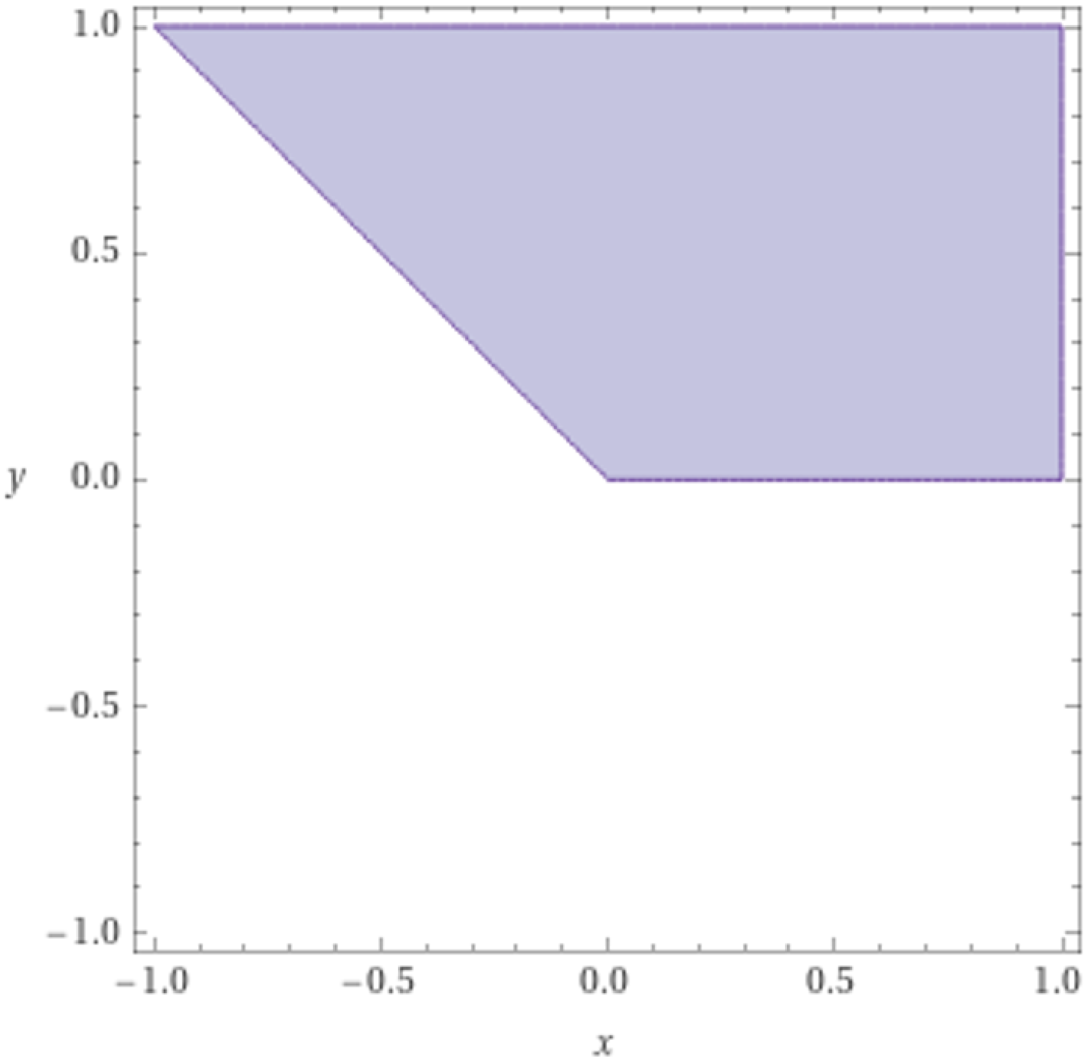
In that context, a measure over X_n is equivalent to a function from a finite domain to R^+; which is equivalent to a point in $(R^+)^{|X_n|}$. So the space of measures over X_n is just the Euclidean space of dimension $|X_n|$. We’re back into linear algebra!

Now geometrical intuition can come to our help. Take Definition 2 of an sa-measure: it is just a point of $(R^+)^{|X_n|+1}$ such that the sum of the negative numbers among its first $|X_n|$ components is less in absolute value than the last component. And an a-measure (from Definition 3) is an sa-measure where every component is non-negative. The sets $M^{sa}(X_n)$ and $M^a(X_n)$ are then respectively the sets of all sa-measures and the sets all a-measures.

We can even visualize them pretty easily (with $|X_n| = 1$):



$M^a(X_n)$



$M^{sa}(X_n)$

There’s one more definition to go through before attacking infradistributions: the definition of the **expectation of some continuous function from X_n to $\mathbb{R}$ by a set of sa-measures** B. This is described by Vanessa and Diffractor as the behavior of f (continuous from X_n to $[0,1]$) according to B. Definition 4 gives $E_B(f)$ as the infinimum of $m(f) + b$ for $(m, b) \in B$. And in our finite case, $m(f) = \sum_{x \in X_n} m(x)f(x)$. So $E_B(f)$ can be rewritten as the infinimum of $\sum_{x \in X_n} m(x)f(x) + b$ for $(m, b) \in B$.

Intuitively, $E_B(f)$ represents the worst expected utility possible over B, where f is the utility function. This fits with our previous discussion of Knightian Uncertainty and Murphy, because we assume that the environment picked (the sa-measure) is the worst possible for us. That is, the one with the worst expected utility.

Geometrically in our finite setting, this is the smallest dot product of a point in $\bar{B}$ with the point of $(R^+)^{|X_n|+1}$ which has for its first $|X_n|$ components the values of f for the corresponding element of X_n, and for its last component 1.

We can finally go to infradistributions: an infradistribution B is just a set of sa-measures satisfying some conditions. I'll now go through them, and try to provide as much intuition as possible.

- **(Condition 1, Nonemptiness)**: $B \neq \emptyset$ This is without a doubt the most complex condition here, but I have faith that you can make sense of it by yourself.
- **(Condition 2, Closure)** $B = \bar{B}$ This condition says that B contains its limit points. Another way to see it is that if you have a set B and you want to make it into an infradistribution, you need to take the closure of B -- add the limit points of B to the set. Why can we do that? Because the definition of expectation uses an infinimum over B, which is basically a minimum over $\bar{B}$. So the expectation, which captures the behavior of utility functions on our set, already takes into account the limit points of B. Adding them will thus maintain all expectations, and not change anything about the behavior of the set.
  Why do it then? There are two ways to see it. First, adding the limit points makes the study of B easier, because closed sets are nicer than generic sets. Among other things, the expectation is now easier to compute, because it doesn't involve taking limits. And second, if B and $\bar{B}$ are distinct, but have the same behavior according to expectations, then they should be collapsed together in some way. (This is Desideratum 2 from Introduction To The Infra-Bayesianism Sequence)
- **(Condition 3, Convexity)** $B = \text{convexHull}(B)$ This is the same kind of condition that the previous one, but instead of adding the limit points, we add the convex combinations of points in B: $(m^{'}, b^{'}) = \lambda(m_1, b_1) + (1-\lambda)(m_2, b_2)$, for $\lambda \in [0,1]$ and $(m_1, b_1), (m_2, b_2) \in M^{sa}(X_n)$. We can add such points because $m^{'}(f) + b^{'} = \lambda(m_1(f) + b_1) + (1-\lambda)(m_2(f) + b_2)$. One of the two components must be smaller or equal to than the other; without loss of generality, let's say it's $m_1(f) + b_1$. Then

  $\lambda(m_1(f) + b_1) + (1-\lambda)(m_2(f) + b_2)$

  $\geq \lambda(m_1(f) + b_1) + (1-\lambda)(m_1(f) + b_1) = m_1(f) + b_1$.

  Hence $(m^{'}, b^{'})$ is not changing the expectation for any f.

- **(Condition 4, Upper-Completion)** $B = B + M^{sa}(X_n)$ Once again, a condition adds points to the set. This one adds all points formed by the sum of an element of B and an element of $M^{sa}(X_n)$. The reason why it's possible is even more intuitive here: we're adding points that have strictly more measure and expected utility, so they don't influence the minimum in the expected value definition, and thus don't change the behavior of the set.
- **(Condition 5, Minimal-positivity)** $B^{min} \subseteq M^{a}(X_n)$ If your set is formed by summing some points with all of $M^{sa}(X_n)$, what is a minimal set of points that would generate your set through this sum? These are the minimal points of B, noted B^{min}. In fact, there is only one such set, because a minimal point cannot be generated from any other point of the set summed with a point of $M^{sa}(X_n)$.

  This condition requires that such minimal points have no negative measure. So there is no element of X_n for which they return a negative number. This is a slightly less straightforward condition to motivate, because it stems from the maths. Basically, a positive measure is just a scaled probability distribution, so it behaves nicely for a lot of purposes. Whereas not all signed measures can be rescaled to probability distribution. So Infra-Bayesianism uses "negative probabilities" for some computations and reasoning (notably to have a stronger upper-closure property), but the end results are really scaled probabilities. This is why requiring minimal points to be a-measures makes sense.
- **(Condition 6a, Minimal-boundedness)** $\exists C$ a compact set such that $B^{min} \subseteq C$ In all honesty, I don't know exactly where this condition matters. The original post says that compactness is used in the proofs, and it is a pretty useful mathematical assumption in general. But I don't know exactly where it pans out, and I'm not convinced that you need to know it for getting the gist of Infra-Bayesianism.
  I'll thus ask you to accept this condition as a mathematical simplification without much philosophical meaning.

  (The actual condition used in the definition of infradistribution is 6b: $f \mapsto E_B(f)$ is uniformly continuous. But it is similarly a mathematical condition without much philosophical weight).
- **(Condition 7, Normalization)** $E_B(1) = 1 \wedge E_B(0) = 0$ This last condition on the other hand has more to tell. Recall that $E_B(f)$ captures the expected utility over B, using f as a utility function. Since by hypothesis the utility of a state is in $[0, 1]$, the function 0 with 0 utility at every state represents the worst-case utility, and the function 1 with utility 1 at every state represents the best-case utility.
  The conditions then simply say that if no state is worth any utility, the expected utility is 0; and if all states have maximal utility, then the expected utility is maximal too, at 1. So our expected utility lies between 0 and 1, and are normalized.

Armed with these conditions, we now understand Definition 7 of $\square X_n$, the set of infradistributions: it contains all the set of sa-measures that satisfy the conditions above.

Another Perspective on Infradistribution: Duality

There is another way to think about infradistributions: as functionals (in this case, applications from functions to R) with specific properties. This duality is crucial in many proofs and to build a better intuition of Infra-Bayesianism.

Given an infradistribution as a set B, how do we get its dual version? Easy: it's the function h defined by $h(f) = E_B(f)$. So the expectation with regard to our set B is the other way to see and define the infradistribution. Theorem 4 states this correspondence, as well as the properties that h gets from being defined in this way through B:

> **Theorem 4, LF-duality, Sets to Functionals:** *If* B *is an infradistribution/bounded infradistribution, then* $h : f \mapsto B(f)$ *is concave, monotone, uniformly continuous/Lipschitz over* $C(X, [0, 1])$*,* $h(0) = 0, h(1) = 1$*, and* $\text{range}(f) \not\subseteq [0, 1] \implies h(f) = -\infty$*.*

Let's look at the properties of h.

- **(Concavity)** h is a concave function -- it's shaped like a hill when seen in few dimensions. This comes simply from the definition of the expectation and some elementary algebraic manipulations. Notably, it doesn't depend on properties of B.
- **(Monotony)** If $f \leq g$ then $h(f) \leq h(g)$. First recall that the usual order in function space with a partial order as codomain is: $f \leq g \iff \forall x : f(x) \leq g(x)$. So intuitively, this means that if the utility by g is greater or equal than the one by f for every outcome in X_n, then the expected utility for g is greater or equal than the expected from f. That makes a lot of sense to me.
  The reason it holds for h comes from the fact that expectation only depends on the minimal points of an infradistribution (Proposition 3, just after the definition of minimal points). And recall that B satisfies Condition 5 as an infradistribution: its minimal points are all a-measures. This matters because that ensures that in $m(f) + b = \sum_{x \in X_n} m(x)f(x) + b$, the m(x) terms are all strictly positive (for $(m, b) \in B^{min}$).

  Since utility functions are also positive, this means that $f \leq g$ entails $h(f) \leq h(g)$ -- because $\forall (m, b) \in B^{min} : m(f) + b \leq m(g) + b$.

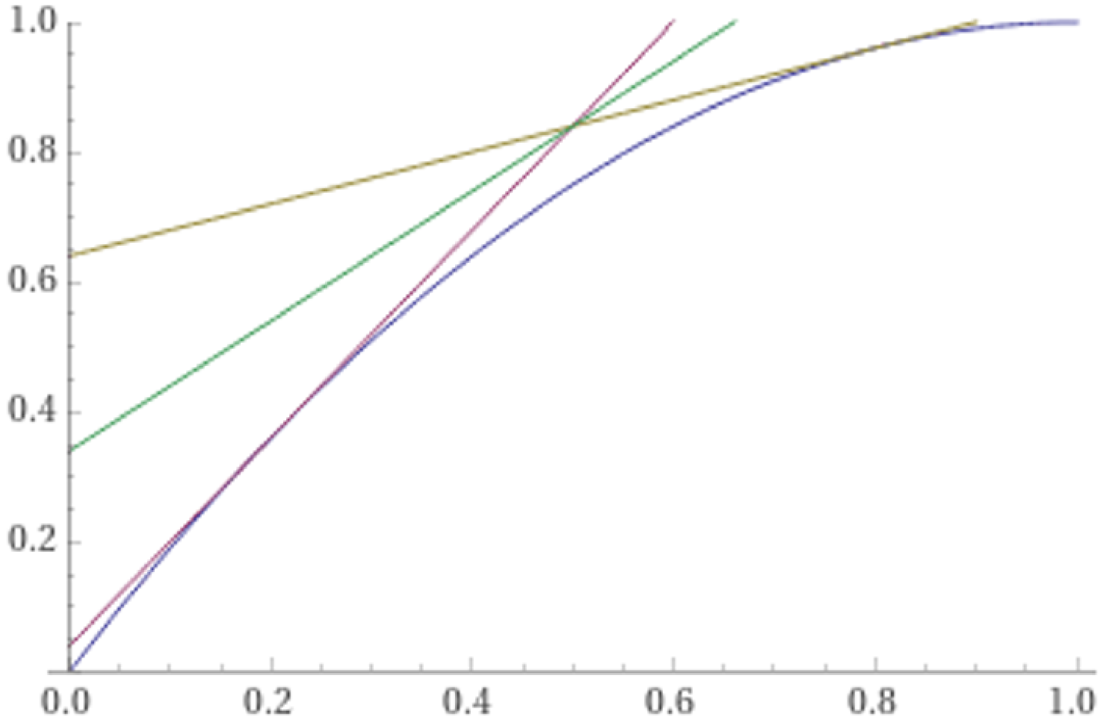
  So this actually depends on B being an infradistribution, specifically the condition that $B^{min} \subseteq M^a(X_n)$**.**

- **(Uniform continuity)** h is uniformly continuous. Well that’s literally Condition 5b for infradistributions.
- **(Normalization)** $h(0) = 0$ and $h(1) = 1$. This is also literally a condition on infradistributions, Condition 7.
- **(Range)** $\mathrm{range}(f) \not\subseteq [0,1] \implies h(f) = -\infty$. So the only functions it makes sense to consider are ones constrained to $[0,1]$. This property follows from Condition 4 on B, upper completion. Indeed, say $x \in X_n$ is such that $f(x) > 1$. Then given any sa-measure in B, we can add to it as many times as we want the sa-measure with -1 measure on x and $b = 1$ (to compensate), thanks to upper completion. Hence we can create sa-measures in B with smaller and smaller $m(f) + b$ forever, which means that taking the infimum, $h(f) = -\infty$.

To summarize, Conditions 4 to 7 for infradistributions as sets do most of the work, while Conditions 1 to 3 are not considered since they just increase the size of the set without changing the expectation (technically Condition 1 is necessary everywhere, but it’s trivial).

Like a healthy relationship, a good duality goes both ways. Hence Theorem 5, which shows how to get an infradistribution as a set from a infradistribution as a functional (satisfying the conditions studied above). The proof of this one is way more involved, which is why I won’t go into it.

That being said, there is a nice way to visualize the set of sa-measures coming from an expectation in the finite dimensional case. Let’s say $|X_n| = 1$. So there is only one outcome x. Let’s say we have an h satisfying all the properties above. Notably, it’s concave. Then the sa-measures of the corresponding infradistribution as a set are all the pairs $(m(x), b)$ such that $m(x)f(x) + b \geq h(f)$ for all f. Visually, any line above h (which is basically a function of $[0,1]$, since f is completely determined by its output for x).



In this plot, h is the blue function, and all other functions correspond to sa-measures in the dual “infradistribution as a set”. This provides a really cool geometrical intuition for some conditions on infradistributions. For example, upper completeness only means that we can add any line to one of our lines/sa-measures, and we’ll still be above h. Or minimal points being a-measures means that they are the tangents of h (like the pink and yellow one on the plot). And it generalizes in higher dimensions, by replacing lines with hyperplanes.

(To be clear, I didn’t come up with this geometric perspective; Diffractor explained it to me during a discussion about the duality).

So infradistributions are both sets of sa-measures and functionals, both satisfying specific conditions. The functional perspective is cleaner for proofs, but I’ll keep with the set perspective in the rest of this post.

Back to Belief Function: Causality and Nirvana

Recall that “nice” belief functions return infradistributions on the outcome set of a policy. This is never stated explicitly in Belief Function and Decision Theory, but follows from the first conditions on belief functions from this section.

Other conditions matter for manipulating belief functions, like consistency and Hausdorff-continuity. But the point of this section isn’t to make you master Belief Function and Decision Theory; it’s to give you a path through it. And the last big idea on the path is Causality, and it’s relation to the Nirvana Trick.

Indeed, if you’ve read the sequence before, you might be surprised by me not mentioning the Nirvana trick already. My reason is that I only understood it correctly after getting causality, and causality requires all the background I layed out already.

The Nirvana Trick: Making Murphy Useful

Recall that we have Knightian Uncertainty over the environments we consider. So instead of maximizing the expected utility over a distribution of environments, we use worst-case reasoning, by assuming the environment is chosen by an adversary Murphy. This is a pretty neat setting, until we consider environments that depend on the policy. This happens notably in Newcomb-like problems (of which Parfit's Hitchhiker is an example), which are an important fighting ground for decision-theories.

Now, it's not so much that representing such environments is impossible; instead, it's that what we think of as environments is usually simpler. Notably, what happens depends only on the action taken by the policy, not on the one it could have taken in other situations. This is also a setting where our intuitions about decisions are notably simpler, because we don't have to think about predictions and causality in their full extent.

The Nirvana trick can be seen as a way to keep this intuition of environments, while still having a dependence of the environment on the policy. It starts with the policy-dependent environment, and then creates one policy-independent environment for each policy, by hard-coding this policy in the parameter slot of the policy-dependent environment. But that doesn't guarantee that the hardcoded policy will match the actual policy. This is where Nirvana appears: if the policy acts differently than the hardcoded policy, it "goes to Nirvana", meaning it gets maximum return (either through an infinite reward at that step or with a reward of 1 for each future step). Murphy, which wants to minimize your utility, will thus never choose an environment where Nirvana can be reached, that is never choose the ones with a different policy in the parameter slot.

To understand better the use of the Nirvana trick, we need to define different kinds of belief functions (called hypotheses) such that adding or removing Nirvana goes from one to the other.

Causality, Pseudocausality and Acausality

The three types of belief functions (called hypotheses) considered in Belief Function and Decision Theory are causal, pseudocausal and acausal. Intuitively, a causal hypothesis corresponds to a set of environments which doesn't depend on the policy; a pseudocausal hypothesis corresponds to a set of environments which depends on the policy in some imperfect way; and an acausal hypothesis corresponds to a set of environments completely and exactly determined by the policy.

Causality can be made formal through the introduction of outcome functions (Definition 15), functions from a policy to a *single* sa-measure on the outcome set of this policy. On the other hand, recall that belief functions return an infradistribution, which is a set of sa-measures on the same set. Compared with the usual Bayesian setting, a belief function returns something analogous to a probability distribution over probability distribution over histories, while an outcome function returns something analogous to a single probability distribution over histories. An outcome function thus plays the role of an environment, which takes in a policy and gives a distribution over the outcomes/histories generated.

There is one additional subtlety about outcome functions that plays a big role in the rest of the formalism. If you look at Definition 15 in Belief Function and Decision Theory, it requires something about the projection of partial policies. The projection mapping (Definition 9) sends a sa-measure over a policy π_1 to a sa-measure over a policy π_2, if π_1 is defined on strictly more histories than π_2 and they agree when they're both defined. Basically, if π_1 extends π_2, we can project back a measure over the outcomes of π_1 to a measure over the outcomes of π_2, by summing the measure of all outcomes of π_1 that share as prefix a given outcome of π_2.

Outcome functions must agree with that, in the sense that the outcome function applied to π_2 must return the projection of what the outcome function returns when applied to π_1. In that sense it's a real environment, because if you extend a policy, it only splits the probability given to each prefix, not moves probability between prefixes.

Causal, pseudocausal and acausal hypotheses are defined through constraints related to the outcome functions corresponding to a belief function. They all share the first 9 Conditions on belief functions given here.

- **Causality** requires Condition C: that for every policy π_{pa} and every sa-measure M from $\theta(\pi_{pa})$ (the application of the belief function to π_{pa}), there is an outcome function f such that the output of f on π_{pa} is M, and the output of f on all other policies is included in the corresponding output of the belief function.
  So for every distribution over history (for a policy), there is an environment sending this policy to this distribution, and any other policy to an accepted distribution over histories (by the belief function). Take all these outcome functions, and you get a set of environments that completely capture the behavior of the belief function.
  Remember that with outcome functions comes a constraint on projections. This constraint does the heavy lifting here: it forces the environments to be policy-independent. This is because it ensures that revealing more actions of the policy (extending it) cannot change what happened before these new actions. Changing an empty policy to the policy that always 1-box in a causal version of transparent Newcomb (so no Omega) doesn't change what is possible to do in a given environment; it merely splits the probability into the extended outcomes.
- **Pseudocausality** requires Condition P. This condition is slightly more complex. First, it requires the belief function to be Nirvana-free: to not consider outcomes leading to Nirvana. So when thinking about pseudocausality, we don't use the Nirvana trick. Starting with a setting without Nirvana, pseudocausality asks that for every sa-measure M from the infradistribution of a given partial policy π, M is also included in the infradistribution for any other policy that generates only outcomes for which M has non-zero measure. So if M only cares about outcomes where both policies agree, it should be either in no infradistribution or in both. This property captures the fact that if two distinct policies don't reveal their difference by taking different actions in a given environment, then this environment should be possible for both or none, but not just one.
  It's a weaker form of policy-independency than Condition C, because it removes the constraint on projection completely -- this definition doesn't even use outcome functions. Note though that there is still some constraint on projection, common to all hypotheses, in the form of Condition 7 on belief functions, consistency. But it's not about policy-dependency, so I won't talk in detail about it.
  Perhaps my biggest initial confusion with Condition P came from the fact that transparent Newcomb with imperfect prediction satisfies it. Intuitively, the environment there should definitely depend on the policy, including on what is done in the other branch. But the trick lies in realizing that imperfect prediction means that no possibility for the transparent box (empty or full) can have probability 0. Thus Condition P doesn't really constrain this problem, because if two policies are different, it will always be revealed by an outcome with non-null measure.

- **Acausality** doesn’t require Condition C or Condition P. It’s for cases where the belief function is so dependent on the policy that pseudocausality fails to hold. Typically, the transparent Newcomb problem with perfect prediction is acausal, because it doesn’t satisfy either Condition C or Condition P.
  To see why, we can focus on Condition P as it’s the weaker condition. Perfect prediction invalidates Condition P because it means for example that the sa-measure giving all measure to the box being empty is in the infradistribution for the policy that 1-box when full and 2-box when empty, yet it isn’t in the infradistribution for the policy that always 2-box. In the latter case, Omega will know that the policy will 2-box on seeing the box full, and thus will make the box full every time.

Perhaps one of the most important results philosophically of Infra-Bayesianism is that one can go from pseudocausality to causality by the Nirvana trick, and from causality to pseudocausality by removing Nirvana (Theorem 3.1). So the first direction basically means that if thinking about policy-dependency fries your brain, you can just add Nirvana, and voilà, everything is policy-independent and causal again. And equivalently, if you have a causal setting with the Nirvana trick, you can remove the trick at the price of only ensuring pseudocausality.

This looks really useful, because in my own experience, non causal situations are really confusing. Having a formal means to convert to a more causal case (at the price of using the Nirvana trick) could thus help in clarifying some issues with decision theory and Newcomb-like problems.

(The same sort of result holds between acausal hypotheses and so-called surcausal hypotheses, but this one requires digging into so many subtle details that I will not present it here.)

Conclusion

Infra-Bayesianism provides a framework for studying learning theory for RL in the context of non-realizability. It is based around infradistribution, sets of distributions with additional data, which satisfy additional conditions for both philosophical and mathematical reasons. Among the applications of Infra-Bayesianism, it can be used to study different decision theory problems in a common framework, and ensure updates which fit with what UDT would do at the beginning of time.

I hope that this post gave you a better idea of Infra-Bayesianism, and whether or not you want to take the time to dig deeper. If you do, I also hope that what I wrote will make navigation a bit easier.

Epistemology of HCH

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

Introduction

HCH is a recursive acronym meaning “Humans consulting HCH”. Coincidentally, It’s also a concept coined by Paul Christiano, central in much of the reasoning around Prosaic AI Alignment. Yet for many, me included, the various ways in which it is used are sometimes confusing.

I believe that the tools of Epistemology and Philosophy of Science can help understand it better, and push further the research around it. So this post doesn’t give yet another explanation of HCH; instead, it asks about the different perspectives we can take on it. These perspectives capture the form of knowledge that HCH is, what it tells us about AI Alignment, and how to expand, judge and interpret this knowledge. I then apply these perspectives to examples of research on HCH, to show the usefulness of the different frames.

Thanks to Joe Collman, Jérémy Perret, Richard Ngo, Evan Hubinger and Paul Christiano for feedback on this post.

Is it a scientific explanation? Is it a model of computation? No, it’s HCH!

HCH was originally defined in Humans Consulting HCH:

> Consider a human Hugh who has access to a question-answering machine. Suppose the machine answers question Q by perfectly imitating how Hugh would answer question Q, *if Hugh had access to the question-answering machine*.
>
> That is, Hugh is able to consult a copy of Hugh, who is able to consult a copy of Hugh, who is able to consult a copy of Hugh…
>
> Let’s call this process HCH, for “Humans Consulting HCH.”

Nowadays, this is actually called weak HCH, after the Strong HCH post which extended the definition. That being said, I’m only interested in perspective about HCH, which includes the questions asked about it and how to answer them. Although the difference between Weak and Strong HCH matters for the answers, the questions and perspective stay the same. I’ll thus use HCH to mean one or the other interchangeably.

The main use of HCH is as an ideal for what a question-answerer aligned with a given human should be like. This in turn serves as the aim of schemes like IDA and Debate. But thinking about HCH entangles many different angles. For example, what can HCH do? One can interpret this question as “What is the power of the HCH scheme?” or “What questions can HCH answer for a given human?” or “Is HCH aligned with the human parametrizing it?” or “Is HCH competitive?”. Each one of these questions requires different assumptions and focus, making it hard to grasp the big picture.

I claim that most of what is studied about HCH can be brought to order if we explicitly use different epistemological perspectives through which to see it. This is close to the paradigms of Kuhn in The Structure of Scientific Revolutions: a framing of the phenomenon studied which explains its most important aspects, how to study them, and what counts as knowledge for this approach. The perspectives I present are both more abstract than paradigms in natural sciences (they’re more epistemological paradigms) and less expansive. Yet I believe the intuition stays the same.

Kuhn writes that paradigms are necessary for what he calls normal science (and which encompass the majority of scientific research), which is solving the puzzles generated by the paradigm. Similarly, the perspectives I propose each put some questions and puzzles in front, and limit the scope of HCH somewhat. Thus not one is supposed to be sufficient; they all have something to bring to the table.

Each of these perspective provide assumptions about HCH:

- What it is
- What are the important questions about it

But before giving these, let’s start with a classical perspective in science that doesn’t work well here.

False start: HCH as explanation of a natural phenomenon

In natural sciences, ideas are often explanations of natural phenomena, like lightning and oxidation. Once armed with such an explanation, the research attempts among other things to check it against previous data of the phenomenon, and to predict new behavior for future experiments.

Of what could HCH be the explanation? In the original post, Paul describes it as

> our best way to precisely specify “a human’s enlightened judgment” [about a question Q]

So the phenomenon is enlightened judgement. Yet this looks more like an ideal than a phenomenon already present in the world.

Indeed, Paul’s Implementing our considered judgement, his first post on the topic as far as I know, presents the similar notion of “considered judgment” as the outcome of a process that doesn’t exist yet.

> To define my *considered judgment* about a question Q, suppose I am told Q and spend a few days trying to answer it. But in addition to all of the normal tools—reasoning, programming, experimentation, conversation—I also have access to a special oracle. I can give this oracle any question Q’, and the oracle will immediately reply with my considered judgment about Q’. And what is my considered judgment about Q’? Well, it’s whatever I would have output if we had performed exactly the same process, starting with Q’ instead of Q.

Seeing HCH as an explanation of enlightened judgment thus fails to be a fruitful epistemological stance, because we don’t have access to considered judgements in the wild to check the explanation.

HCH as philosophical abstraction

If enlightened judgment isn’t a phenomenon already existing in the world, intuitions nonetheless exist about what it means. For example, it feels like an enlightened judgment should depend on many different perspectives on the problem instead of only on the most obvious one. Or that such judgment shouldn’t change without additional information.

This leads to the perspective of HCH as a philosophical abstraction of the fuzzy intuitions around enlightened judgment (on a question Q). The aim of such an abstraction is to capture the intuitions in a clean and useful way. We’ll see a bit later for what it should be useful for.

How should we judge HCH as a philosophical abstraction of enlightened judgement? One possible approach is inspired by Inference to the Best Explanation with regard to intuitions, as presented by Vanessa in her research agenda:

> Although I do not claim a fully general solution to metaphilosophy, I think that, pragmatically, a quasiscientific approach is possible. In science, we prefer theories that are (i) simple (Occam's razor) and (ii) fit the empirical data. We also test theories by gathering further empirical data. In philosophy, we can likewise prefer theories that are (i) simple and (ii) fit intuition in situations where intuition feels reliable (i.e. situations that are simple, familiar or received considerable analysis and reflection). We can also test theories by applying them to new situations and trying to see whether the answer becomes intuitive after sufficient reflection.

In this perspective, the intuitions mentioned above play the role of experimental data in natural sciences. We then want an abstraction that fits this data in the most common and obvious cases, while staying as simple as possible.

What if the abstraction only fit some intuitions but not others? Here we can take note from explanations in natural sciences. These don’t generally explain everything about a phenomenon -- but they have to explain what is deemed the most important and/or fundamental about it. And here the notion of “importance” comes from the application of the abstraction. We want to use enlightened judgement to solve the obvious question: “How to align an AI with what we truly want?” (Competitiveness matters too, but it makes much more sense from the next perspective below)

Enlightened judgement about a question serves as a proxy for “what we truly want” in the context of a question-answerer. It’s important to note that this perspective doesn’t look for the one true philosophical abstraction of enlightened judgment; instead it aims at engineering the most useful abstraction for the problem at hand -- aligning a question-answerer.

In summary, this perspective implies the following assumptions about HCH.

- **(Identity)** HCH is a philosophical abstraction of the concept of enlightened judgment for the goal of aligning a question-answerer

- **(Important Questions)** These includes pinpointing of the intuitions behind enlightened judgment, weighting their relevance to aligning a question-answerer, and check that HCH follows them (either through positive arguments or by looking for counterexamples)

HCH as an intermediary alignment scheme

Finding the right words for this one is a bit awkward. Yes, HCH isn't an alignment scheme proper, in that it doesn't really tell us how to align an AI. On the other hand, it goes beyond what is expected of a philosophical abstraction, by giving a lot of details about how to produce something satisfying the abstraction.

Comparing HCH with another proposed philosophical abstraction of enlightened judgement makes this clear. Let's look at Coherent Extrapolated Volition (CEV), which specify "what someone truly wants" as what they would want if they had all the facts available, had the time to consider all options, and knew enough about themselves and their own processes to catch biases and internal issues. By itself, CEV provides a clarified target to anyone trying to capture someone's enlightened judgement. But it doesn't say anything about how an AI doing that should be built. Whereas HCH provides a structured answer, and tells you that the solution is to get as closed as possible to that ideal answer.

So despite not being concrete enough to pass as an alignment scheme, HCH does lie at an intermediary level between pure philosophical abstractions (like CEV) and concrete alignment schemes (like IDA and Debate).

So what are the problems this perspective focuses on? As expected from being closer to running code, they are geared towards practical solutions:

- How much can HCH be approximated?
- How competitive is HCH (and its approximations)?

That is, this perspective cares about the realization of HCH, assuming it is what we want. It's not really the place to wonder how aligned HCH is; knowing it is, we want to see how to get it right in a concrete program, and whether it costs too much to build.

In summary, this perspective implies the following assumptions about HCH.

- **(Identity)** HCH is an intermediary alignment scheme for a question-answerer.
- **(Important Questions)** Anything related to the realization of HCH as a program: approximability, competitiveness, limits in terms of expressiveness and power.

HCH as a model of computation

Obvious analogies exist between HCH and various models of computations like Turing Machines: both give a framework for solving a category of problems -- questions on one side and computable problems on the other. HCH looks like it gives us a system on which to write programs for question answering, by specifying what the human should do.

Yet one difficulty stares us in the face: the H in HCH. A model of computation, like Turing Machines, is a formal construct on which one can prove questions of computability and complexity, among others. But HCH depends on a human at almost every step in the recursion, making it impossible to specify formally (even after dealing with the subtleties of infinite recursion).

Even if one uses the human as a black box, as Paul does, the behavior of HCH depends on the guarantees of this black box, which are stated as “cooperative”, “intelligent”, “reasonable”. Arguably, formalizing these is just as hard as formalizing good judgment or human values, and so proves incredibly hard.

Still, seeing HCH through the perspective of models of computation has value. It allows the leveraging of results from theoretical computer science to get an idea of the sort of problems that HCH could solve. In some sense, it’s more OCO -- as in “Oracles Consulting OCO”.

Knowledge about HCH as a model of computations is thus relatively analogous to knowledge for Turing Machines:

- Upper bounds for computability and complexity (algorithms)
- Lower bounds for computability and complexity (impossibility results)
- Requirements or structural constraints to solve a given problem

In summary, this perspective implies the following assumptions about HCH.

- **(Identity)** HCH is a model of computation where the human is either a Turing Machine or an oracle satisfying some properties.
- **(Important Questions)** These include what can be computed on this model, at which cost, and following which algorithm. But anything that is ordinarily studied in theoretical computer science goes: simulation between this model and others, comparison in expressivity,...

Applications of perspectives on HCH

Armed with our varied perspective on HCH, we can now put in context different strands of research about HCH that appear incompatible at first glance, and to judge them with the appropriate standards. The point is not that these arguments are correct; we just want to make them clear. After all, it’s not because someone makes sense that they’re right (especially if they’re disagreeing with you).

Here are three examples from recent works, following the three perspectives on HCH from the previous section. Yet keep in mind that most if not all research on this question draws from more than one perspective -- I just point to the most prevalent.

Assumptions about H to have an aligned Question-Answerer

In a recent post (that will probably soon have a followup focused on HCH), Joe Collman introduced the Question-Ignoring Argument (QIA) in Debate:

- For a consequentialist human judge, the implicit question in debate is always “Which decision would be best for the world according to my values?”.
- Once the stakes are high enough, the judge will act in response to this implicit question, rather than on specific instructions to pick the best answer to the debate question.
- With optimal play, the stakes are always high: the value of timely optimal information is huge.
- The “best for the world” output will usually be unrelated to the question: there’s no reason to expect the most valuable information to be significantly influenced by the question just asked.
- The judge is likely to be persuaded to decide for the “best for the world” information on the basis of the training signal it sends: we already have the high-value information presented in this debate, even if it doesn’t win. The important consequences concern future answers

The gist is that if the human judge is assumed to have what we would consider outstanding qualities (being a consequentialist that wants to do what’s best for the world), then there is an incentive for the debaters to give an answer to a crucially important question (like a cure for a certain type of cancer) instead of answering the probably very specific question asked. So there is a sense in which the judge having traits we intuitively want makes it harder (and maybe impossible) for the system to be a question-answerer, even if it was the point of the training.

Applying the same reasoning to HCH gives a similar result: if H is either an altruistic human or a model of such a human, it might answer more important questions instead of the one it was asked.

Regardless of whether this line of thinking holds, it provides a very good example of taking HCH as a philosophical abstraction and investigating how much it fits the intuitions for enlightened judgement. Here the intuition is that the enlightened judgement about a question is an enlightened answer to this question, and not just a very important and useful (but probably irrelevant to the question) information.

Experimental work on Factored Cognition

Ought has been exploring Factored Cognition through experiments for years now. For example, their latest report studies the result of an experiment for evaluating claims about movie review by seeing only one step of the argument.

Such work indirectly studies the question of the competitiveness of HCH. In a sense, the different Factored Cognition hypotheses are all about what HCH can do. This is crucial if we aim to align question-answers by means of approximating HCH.

The Ought experiments attempt to build a real-world version of (necessarily bounded) HCH and to see what it can do. They thus place themselves in the perspective of HCH as an intermediary alignment scheme, focusing on how competitive various approximations of it are. Knowing this helps us understand that we shouldn’t judge these experiments through what they say about HCH alignment for example, because their perspective takes it for granted.

HCH as Bounded Reflective Oracle

In Relating HCH and Logical Induction, Abram Demski casts HCH as a Bounded Reflective Oracle (BRO), a type of probabilistic oracle Turing Machine which deals with diagonalization and self-referential issue to answers questions about what an oracle would do when given access to itself (the original idea is of Reflective Oracle -- the post linked above introduces the boundedness and the link with Logical Induction.) This reframing of HCH allows a more formal comparison with Logical Induction, and the different guarantees that they propose.

The lack of consideration of the human makes this post confusing when you think first of HCH as a philosophical abstraction of enlightened judgement, or as an intermediary alignment scheme. Yet when considered through the perspective of HCH as a model of computation, this makes much more sense. The point is to get an understanding of what HCH actually does when it computes, leveraging the tools of theoretical computer science for this purpose.

And once that is clear, the relevance to the other perspectives starts to appear. For example, Abram talks about different guarantees of rationality satisfied in Logical Induction, and why there is no reason to believe that HCH will satisfy them by default. On the other hand, that raises the question of what is the impact of the human on this:

> It would be very interesting if some assumptions about the human (EG, the assumption that human deliberation eventually notices and rectifies any efficiently computable Dutch-book of the HCH) coud guarantee trust properties for the combined notion of amplification, along the lines of the self-trust properties of logical induction.

Conclusion

HCH is still complex to study. But I presented multiple perspectives that help clarify most discussions on the subject: as a philosophical abstraction, as an intermediary alignment scheme, and as a model of computation.

Nothing guarantees that these are the only fruitful perspectives on HCH. Moreover, these might be less useful than imagined, or misguided in some ways. Yet I’m convinced, and I hope you’re more open to this idea after reading this post, that explicit thinking about which epistemic perspectives to take on an idea like HCH matters to AI Alignment. This is one way we make sense of our common work, both for pushing more research and for teaching the newcomers to the field.

Epistemological Framing for AI Alignment Research

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

Introduction

You open the Alignment Forum one day, and a new post stares at you. By sheer luck you have some time, so you actually read it. And then you ask yourself the eternal question: how does this fit with the rest of the field? If you're like me, your best guess comes from looking at the author and some keywords: this usually links the post with one of the various "schools" of AI Alignment. These tend to be affiliated with a specific researcher or lab -- there's Paul Christiano's kind of research, MIRI's embedded agency, and various other approaches and agendas. Yet this is a pretty weak understanding of the place of new research.

In other fields, for example Complexity Theory, you don't really need to know who wrote the paper. It usually shows a result from one of a few types (lower bound, completeness for a class, algorithm,...), and your basic training in the field armed you with mental tools to interpret results of this type. You know the big picture of the field (defining and separating complexity classes), and how types of results are linked with it. Chances are that the authors themselves called on these mental tools to justify the value of their research.

In the words of Thomas S. Kuhn, Complexity Theory is paradigmatic and AI Alignment isn't. Paradigms, popularized in Kuhn's The Structure of Scientific Revolutions, capture shared assumptions on theories, interesting problems, and evaluation of solutions. They are tremendously useful to foster normal science, the puzzle-solving activity of scientists; the paradigm carves out the puzzles. Being paradigmatic also makes it easier to distinguish what's considered valuable for the field and what isn't, as well as how it all fits together.

This list of benefit logically pushed multiple people to argue that we should make AI Alignment paradigmatic.

I disagree. Or to be more accurate, I agree that we should have paradigms in the field, but I think that they should be part of a bigger epistemological structure. Indeed, a naive search for a paradigm either results in a natural science-like paradigm, that put too little emphasis on applications and usefulness, or in a premature constraint on the problem we're trying to solve.

This post instead proposes a framing of AI Alignment research which has a place for paradigms, but isn't reduced to them. I start by stating this framing, along with multiple examples in each of its categories. I then go back to the two failure modes of naive paradigmatism I mentioned above. Finally, I detail how I intend to falsify the usefulness of this framing through a current project to review important AF posts.

Thanks to Joe Collman, Jérémy Perret, Evan Hubinger, Rohin Shah, Alex Turner and John S. Wentworth for feedback on this post.

The Framing

Let’s start by asking ourselves the different sort of progress one could make in AI Alignment. I see three categories in broad strokes (I’ll give examples in a minute).

- Defining the terms of the problem
- Exploring these definitions
- Solving the now well-defined problem

I expect the first and third to be quite intuitive -- define the problem and solve it. On the other hand, the second might feel redundant. If we defined the problem, the only thing left is to solve it, right?

Not in a world without logical omniscience. Indeed, the definitions we’re looking for in AI Alignment are merely structures and premises; they don’t give all their consequences for free. Some work is needed to understand their implications.

Let’s get slightly less abstract, and try to state the problem of AI Alignment: “Make AIs well-behaved". Here “AIs” and “well-behaved” are intentionally vague; they stand for “AI-related systems we will end up building” and “what we actually want them to do”, respectively. So I’m just saying that AI Alignment aims to make the AIs we build do as we wish.

What happens when we try to carve research on this abstract problem along the three categories defined above?

- **Research on the “AIs” part**
    - **(Defining)** Clarify what “AI-related systems we will end up building” means. This basically amounts to making a paradigm for studying the AIs we will most probably build in the future.
      Note that such a paradigm is reminiscent of the ones in natural sciences, since it studies an actual physical phenomenon (the building of AIs and what they do, as it is done).
      Examples include:
        - Timelines research, like Daniel Kokotajlo’s posts
    - **(Exploring)** Assuming a paradigm (most probably deep learning these days), this is normal science done within this paradigm, that helps understanding aspects of it deemed relevant for AI Alignment.
      Examples (in the paradigm of deep learning) include:
        - Interpretability work, like the circuit work done by the Clarify team at OpenAI.
        - Work on understanding how training works, like this recent work on SGD
- **Research on the “well-behaved” part**
    - **(Defining)** Clarifying what “what we actually want them to do” means. So building a paradigm that makes clear what the end-goals of alignment are. In general, I expect a global shared paradigm here too, with individual researchers championing specific properties among all the ones promoted by the paradigm.
      Note that such a paradigm is reminiscent of the ones in theoretical computer science, since it studies a philosophical abstraction in a formal or semi-formal way.
      Examples include:

        - Defining Coherent Extrapolated Volition as an abstraction of what we would truly want upon reflection.
        - Defining HCH as an abstraction of considered judgment
        - Defining and arguing about corrigibility
        - Defining the properties expected of good embedded agents.
        - Defining catastrophic consequences through attainable utility.
    - **(Exploring)** Assuming a paradigm (or at least some part of the paradigm focused on a specific property), normal science done in extending and analyzing this property.
      Examples include:
        - Assuming "well-behaved" includes following considered judgement, works on exploring HCH, like these two posts.
        - Assuming "well-behaved" includes being a good embedded agent, works on exploring embedded agency, like the papers and posts referenced in the Embedded Agency sequence.
- **(Solving)** Assuming a paradigm for "AIs" and a paradigm for "well-behaved", research on actually solving the problem. This category is probably the most straightforward, as it includes most of what we intuitively expect in AI Alignment research: proposition for alignment schemes, impossibility results, critics of schemes,...
  Examples include:
    - Assuming "AIs" means "Deep Learning models for question answering" and "well-behaved" means "following HCH", IDA is a proposed solution
    - Assuming "AIs" means "DeepRL systems" and "well-behaved" means "coherent with observed human behavior", an impossibility result is the well-known paper on Occam Razor's and IRL by Stuart Armstrong and Sören Mindermann.
    - Assuming "AIs" means "Embedded Agents" and "well-behaved" means "deals with logical uncertainty in a reasonable way", logical inductors are a proposed solution.

Note that this framing points towards some of the same ideas that Rohin's threat models (I wasn't aware of them before Rohin's pointer in an email). Basically, Rohin argues that a model on which to do AI Alignment research should include both a development model (what AI will look like) and a risk model (how it will fail). His issue with some previous work lies in only filling one of these models, and not both. In my framing, this amounts to requiring that work in the Solving category comes with both a model/paradigm of what "AIs" means and a model/paradigm of what "well-behaved" means. That fits with my framing. On the difference side, Rohin focuses on "what goes wrong" (his risk model), whereas I focus on "what we want".

Going back to the framing, let's be very clear on what I'm **not** saying.

I'm not saying that every post or paper falls within exactly one of these categories. The Logical Induction paper for example both defines a criterion for the part of "well-behaved" related to embedded logical uncertainty, but also provides logical inductors to show that it's possible to satisfy it. Yet I think it's generally easy to separate the different contributions to make clear what falls into which category. And I believe such explicit separation helps tremendously when learning the field.

I'm not saying that these categories are independent. It's obvious that the "solution" category depends on the other two; but one can also argue that there are dependencies between studying what "AIs" means and studying what "well-behaved" means. For example, inner alignment only really makes sense in a setting where AIs are learned models through some sort of local optimization process -- hence this part of “well-behaved” requires a specific form to the definition of “AIs”. This isn’t really a problem, though.

I’m not saying that every post or paper falls within at least one category. Some work that we count as AI Alignment don’t really fall in any of my categories. The foremost example that I have in mind is John’s research on Abstraction. In a way, that is expected: this research is of a more general idea. It impacts some categories (like what “well-behaved” means), but is more a fundamental building block. Still, pointing to the categories that this research applies might help make it feel more relevant to AI Alignment.

I’m not saying that we need to fully solve what we mean by “AIs” and “well-behaved” before working on solutions. Of course work on solutions can already proceed quite usefully. What I’m arguing for instead is that basically any work on solutions assumes (implicitly or explicitly) some sort of partial answer to what “AIs” and “well-behaved” means. And that by stating it out loud, the authors would help the understanding of their work within the field.

I’m not saying that this is the only reasonable and meaningful framing of AI Alignment research. Obviously, this is but one way to categorize the research. We already saw that it isn’t as clean as we might want. Nonetheless, I’m convinced that using it will help make the field clearer to current researchers and newcomers alike.

In essence, this framing serves as a lens on the field. I believe that using it systematically (as readers when interpreting a work and as author when presenting our work) would help quite a lot, but that doesn’t mean it should be the only lens ever used.

Why not a single paradigm?

I promised in the introduction that I would explain why I believe my framing is more adequate than a single paradigm. This is because I only see two straightforward ways of compressing AI Alignment into a single paradigm: make it a paradigm about a fundamental abstraction (like agency) that once completely understood should make a solution obvious; or make it a paradigm about a definition of the problem (what “AIs” and “well-behaved” means). Both come with issues that make them undesirable.

Abstraction Paradigm

Paradigms historically come from natural sciences, as perspectives or explanations of phenomena such as electricity. A paradigm provides an underlying theory about the phenomenon, expresses the well-defined questions one can ask about it, and what would count as a successful solution of these questions.

We can also find paradigms about abstractions, for example in theoretical computer science. The current paradigm about computability is captured by the Church-Turing thesis, which claims that everything that can be physically computed can be computed by a Turing Machine. The “explanation” for what computation means is the Turing Machine, and all its equivalent models. Hence studying computability within this paradigm hinges on studying what Turing Machines can compute, as well as other models equivalent to TMs or weaker (This overlooks the sort of research done by mathematicians studying recursion theory, like Turing degrees; but as far as I know, these are of limited interest to theoretical computer scientists).

So a paradigm makes a lot of sense when applied to the study of a phenomenon or an abstraction. Now, AI Alignment is neither; it’s instead the search for the solution of a specific problem. But natural sciences and computer science have been historically pretty good at providing tools that make solving complex problems straightforward. Why couldn’t the same be true for AI Alignment?

Let's look at a potential candidate. An abstraction presented as the key to AI Alignment by multiple people is agency. According to this view, if we had a complete understanding of agency, we wouldn’t find the problem of aligning AI difficult anymore. Thus maybe a paradigm giving an explanation of agency, and laying out the main puzzles following from this explanation, would be a good paradigm of AI Alignment.

Despite agreeing with the value of such work, I disagree with the legitimacy of making it the sole paradigm of AI Alignment. Even if understanding completely something like agency would basically solve the problem, how long will it take (if it is ever reached)? Historical examples in both natural sciences and computer science show that the original paradigm of a field isn’t usually adapted to tackle questions deemed fundamental by later paradigms. And this progress of paradigms takes decades in the best of cases, and centuries in the worst!

With the risk of short timelines, we can’t reasonably decide that this is the only basket to put our research eggs.

That being said, this paradigmatic approach has a place in my framing, about what “well-behaved” means. The difference is that once a paradigm is chosen, work can proceed in it while other researchers attempt to solve the problem for the current paradigm. There’s thus a back and forth between the work within the paradigm and its main application.

Problem Paradigm

If we stretch a bit the term, we can call paradigm the assumptions about what “AIs” and “well-behaved”. Then becoming paradigmatic would mean fixing the assumption and forcing all the work to go within this context.

That would be great, if only we could already be sure about what assumptions to use. But in the current state of the field, a lot more work is needed (especially for the “well-behaved” part) before anyone can reasonably decide to focus all research on a single such paradigm.

This form of paradigm thus suffers from the opposite problems than the previous one: it fails to value the research on the term of the problems, just to have a well-defined setting on which to make progress. Progress towards what? Who knows…

Here too, this approach has a place in my framing. Specifically, every work on the Solving category exists within such a paradigm. The difference is that I allow multiple paradigms to coexist, as well as the research on the assumptions behind this paradigm, allowing a saner epistemological process.

Where do we go from here?

Multiple voices in AI Alignment push for making the field more paradigmatic. I argue that doing this naïvely isn’t what we want: it either removes the push towards application and solutions, or fixes the term of the problem even though we are still so uncertain. I propose instead that we should think about research according to different parts of the statement "Make AIs well-behaved": research about what “AIs” we're talking about, research on what we mean by “well-behaved”, and based on answers to the two previous questions, actually try to solve the clarified problem.

I believe I argued reasonably enough for you to not dismiss the idea immediately. Nonetheless, this post is hardly sufficient to show the value of adopting this framing at the level of the whole research community.

One way I hope to falsify this proposition is through a project to review many posts on the AF to see what makes a good review, done with Joe Collman and Jérémy Perret. We plan on trying to use this lens when doing the reviews, to see if it clarifies anything. Such an experiment thus relies on us reviewing both posts that fit quite well the framing, and ones that don’t. If you have any recommendation, I wrote a post some time ago where you can give suggestions for the review.