Will AGI Learn Alien Concepts?

I care about the problem of what representations of objects an AI will learn for two reasons.

First because an agent's goals or values are specified in terms of its world model. This is true even if we are talking about very simple agents. Take a simple paperclip maximizer. Its utility function is defined as the number of paperclips in the world. This means its model of the world must include paperclips. Now I think we are very much in the dark about what it would actually mean for the agent's world model to successfully latch onto paper clips at all, much less in a way which is reflexively stable (i.e., not subject to ontological crisis), such that it could orient its actions on the basis of them. We do not understand how agents learn representations of the world (and I'm not even sure we understand what this capacity actually amounts to, more below). We also don't have any clear idea of how to load specific representations into an AI. We don't know how to point their actions to things picked out by those representations robustly. I would like to understand this better.

That is very general though. I also wonder about whether or not the representations an AGI will learn will be alien. Will it learn concepts that are like the ones we employ, or will it think in very strange ways that are incommensurable with our ordinary ways of thinking?

I worry about this because if we hope to be able to understand agents using interpretability techniques it will help if the agent uses representations which are not totally alien. If an AGI learns and then takes actions on the basis of concepts which are more advanced than the ones we employ then this will make it harder to interpret. For example, if it is thinking in terms of exotic physics to make predictions about the world, we will probably have a hard time understanding why it is doing the things it is doing for the same reason that if a contemporary physicist showed up in 500 AD with a bunch of electronics they would probably have a very hard time explaining how those things works.

Even if it could *explain* the things it was doing on the basis of alien representations, that would add an extra point of failure in our safety setups. It would be easy for the system to lie or be misleading. We would be throwing our interpretability tools at something even more incomprehensible than our current systems. And even if we could make some headway to see through outright deception and are using interpretability to directly inspect its internal reasoning processes, if its reasoning is done in very alien or strange ways it would also make it easy for a system to shield its thoughts from us.

I would also know how likely it is that an agent might suddenly develop novel conceptual structures that might suddenly break an otherwise functional ELK translator's head.

What Even Is An Agent's Ontology?

I want to lay out how I think of a few different questions related to interpretability and alignment. Some of my terminology is borrowed and some of my terminology diverges from how some alignment people I have seen have written about it. I want to introduce three classes of questions we can ask about an agent's internal ontology.

The *ontology specification* problem is "what ontology does this agent have?" What is the basic structure in terms of which the agent thinks about the world? Is it thinking in terms of atoms and the void, in terms of classical physics, in terms of quantum mechanics, in terms of everyday medium size objects (tables, chairs, people etc)? There is no reason to suppose that an agent must have exactly one ontology for the same reason that humans do not. We just want to know what the agent's world model is and includes.

Some sub-questions here include:

> (1) can we find specific things in our ontology in the agent's ontology? For example, we have the concept of a human being. Does this exist anywhere in the AI's ontology?
>
> (2) For specific planning or reasoning episodes, what abstractions is it employing to generate its answer? If we ask it a question we might want to know what it is thinking in terms of when it generates its answer. So if we ask it to predict some planetary movements, what kind of physics knowledge is it employing? If it predicts what a person will do, is it employing some kind of weird model or is it thinking in terms of the kind of abstractions we use to predict human behavior? If we are asking it what it knows, is it using a direct translator or a human simulator? Are the answers it is using conditioned on whether or not it believes it is in a training environment?

Normally we solve the ontology specification problem from the outside, for humans at least. We ask other human beings what they believe and derive an ontology for them from the content of their answers. This method has three problems though: it requires that the other person is honest, has a language, and shares an ontology with you, at least in rough outline.

If we aren't deriving an ontology from conversation then we resort to the pragmatic attribution of concepts and beliefs (the ingredients of an ontology) pragmatically. For less complicated agents that can't speak (e.g, animals) if we are attributing to them an ontology we only have the second strategy available, and so there is often a lot of ambiguity (does my dog really love me? Does it even have a representation of a human being?).

Advanced interpretability for AGI means we could do this from the inside. There is the potential that we could actually understand just by looking at the system's internal representations what sort of ontology it has. This might not work because it is incoherent conceptually, or this might be an engineering feat that is too hard or easily defeated by uncooperative systems.

The *ontology identification* problem is "how can we get an agent with a specific ontology?". This might mean generally, but it might mean specifically. We might want to have an agent that has a particular general way of thinking about the world. It could be stuck to a specific kind of regional ontology. For example we might want its ontology to only include fundamental physics so that it is more difficult to model human psychology easily and use that knowledge to manipulate human beings. Or we might wonder how we can get the model's ontology to include *specific things* like paper clips or people for example.

If the natural abstractions hypothesis is true, this might be easy. There are basic features of the world which lead almost all cognitive architectures to converge to specific ontologies. Or it might be the case that learning a natural language *eo ipso* means learning the ontology(ies) that the language specifies so any agent which successfully learned a human language would ipso facto have acquired the ontology specified in the concepts of that language (I think that there are aspects of this view that are partly correct, but it isn't the concern here).

Related sub-questions here include: how can an agent find the things picked out by its ontology in the world? What are the conditions under which this breaks down? How does an agent's world model shift as it enters new domains other than its training set? Can we "lock" an agent's ontology (especially without otherwise breaking its learning ability - i.e, is it the case that any general learning ability will mean that your ontology is unstable because you can't reliably update beliefs if you can't generate new conceptual structures to support, or more minimally that this puts a serious cap on your learning ability?).

There is also the dynamic of an agent's world model. When talking about shifts in an agent's world model, I call an ontic shift coming to believe in the existence or nonexistence of some particular entity, like if I come to believe that Fido is a dog and he exists. This is just a way of saying acquiring a belief which involves material existential quantification. The more important shift I call an ontological shift. That means coming to believe or not believe in an entire class of entities. For example, you might come to think there really are no such things as dogs; the category is an illusion. Or you might radically reshape what you take the basic structure of those entities to be like going from thinking of entities as the union of essence and accident to thinking of them in terms of brute quantitative substance. Both are the kind of shifts which could make a big change to an entity's behavior, although only the second kind of shift touches on the ontology problem in particular.

For example of an ontological shift and the reason we might want to know how to lock an agent's ontology so it doesn't change the ways in which it thinks about the world is that we might know how to specify useful safety properties like corrigibility in a system. But then the agent might suffer an ontological reshaping where it stops believing that there really are such things as human beings and then does something really weird (and potentially catastrophic). That is obviously undesirable. I don't know how much of a problem this will end up being in practice, but it could be quite significant.

Ontology loading and ontology locking are the ontology engineering questions which have the most *direct* significance for alignment. If we knew how to load an agent's ontology it might make it easier to direct its actions towards specific entities or values in the world.

The last problem is what I call the *ontology explanation* problem. This question is "what does it mean for an agent to have an ontology". When we say that agent A has ontology X, what are we actually saying of the agent? This is the part of the problem that relates to simple and straightforward questions like: what are concepts, what are abstractions, what is knowledge, what is truth, what is meaning, what is thinking?

Addressing the Actual Question

With some terminological ground covered we can now move to a more direct consideration. We are here considering some aspects of the ontology identification problem. What kind of factors push towards agents to have non-human ways of forming abstractions of the world? I'll now lay out some general factors that push towards alien abstractions:

(1) *Reality Underdetermines Concepts.* There might be many ways of carving up reality, and the choice of concepts you use to do this is heavily influenced by pragmatic considerations or the structure of your cognitive architecture. For example, things that are very salient to human beings in our environment based on either our needs or the structure of our brain that enter into our languages and classifications might not be very salient to an alien observer who landed on our planet and began making concepts to describe the world around it. Many of our abstractions might be highly parochial. I think this is probably something of a mix. There are some concepts of which I think this is true, and some where this is probably less true, more below.

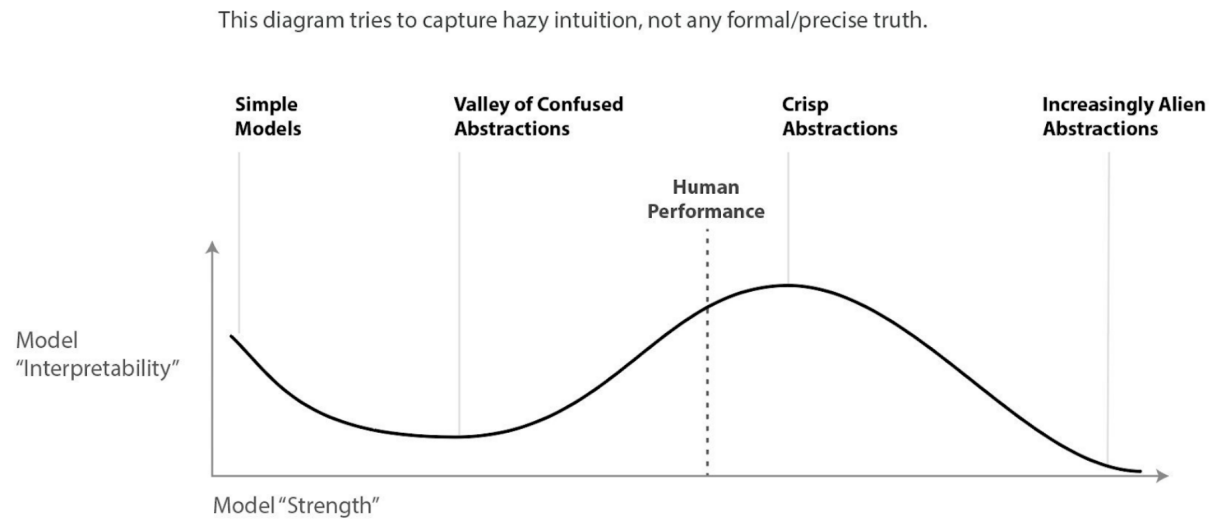
But if this is importantly true then an AGI might interpret reason about the using quite different concepts than the one humans do because if there are many ways of constructing concepts to understand things, just by the math we would expect that we won't get human ways of understanding unless we take efforts to make agents that use them.

(2) *More Useful Abstractions Are Alien.* There might be more useful abstractions than the ones humans employ, such that if we deploy an AI that is aimed at some particular task, it will invent/learn alien abstractions in order to do the task well. For example, if it is trying to have a high degree of predictive accuracy to answer difficult scientific questions, it seems likely it could develop novel physical or mathematical structures in order to answer such questions.

There is plausibly a computation/speed penalty here. It requires an investment of computing resources to develop novel abstractions and models that use them. It may also be more computationally expensive to run those novel models (think of the difference between calculating the trajectory of a cannonball using classical mechanics vs using a model which goes all the way to quantum mechanics). Conversely, it might be the case that once you have more advanced abstractions, many problems become dramatically computationally easier. Think of learning some more advanced mathematical concepts which make certain problems much easier or allow you to immediately pattern match with situations like: “novel situation X is actually an instance of advanced structure Y which I can now immediately solve”. But the more advanced an AGI is the less a concern these computational penalties will be.

(3) *Ground Truth is Very Alien.* Opposite to (1) and building on (2) it might be the case that the ground truth of the world is very alien and different from the world model specified in the concepts we deploy. If we think that ground truth exerts a strong pressure on the representations an agent learns in its model of the world, then in this case we should expect that as AGI gets more advanced and learns to model the world better, its concepts will increasingly diverge from ours until they become very alien.

(2) and (3) are similar but slightly different. It might be that the idea of “ground truth” for the universe doesn’t make sense, but there is still a difference to be made for the usefulness of abstractions and so we will get a pull towards more alien abstractions. Or of course it could be that because ground truth is very alien that more alien abstractions are more useful. Now Combining (2) and (3) gives us reason to expect the shape of an AGI’s abstractions to look something like this (from a chart by Chris Olah) :



I now want to move through some kinds of concepts and think about some of the factors that might lead to alien/non-alien abstractions, and then I'll talk about John Wentworth's natural abstractions hypothesis research agenda which points to some similar questions.

Empirical/Perceptual Concepts

By an empirical or perceptual concept, I mean the concepts that directly apply to perceptual objects in the environment. I have in mind primarily sortal concepts that specify entities exemplifying kinds (tables, chairs, trees, beds, people, food, etc), the concepts of the simple perceptible qualities which are attributed to them (round, straight, colored, etc), and the representations that specify unique perceptual individuals (*this person*, *this* diamond, etc).

I think that the chances that advanced AGI will advance to alien abstractions for what is specified in these concepts is unlikely. I want to note that this is a very different claim than the claim that any given cognitive architecture will learn human-like representations. What I mean is that, given that the agent has learned how to robustly pick out the thing marked off by the human concept, will its learned representation become more alien as it becomes more advanced?

For example, if a weak system can learn to classify dogs vs non-dogs by confused abstractions and a more advanced system learns to pick up on features that are similar to what humans learn (this seems to be true of a lot of current image recognition schemes), what representations will its more advanced successor learn? I think it plausibly the case that at the limit, there might not be a better way to predict classificatory membership in one of these kinds than by latching onto the same features that humans use to. For these kinds of concepts, I think that dip in the chart above is unlikely.

Theoretical/mathematical Concepts

When it comes to theoretical concepts, those that we use to understand the basic structure of reality, I think alien concepts are very likely, for reasons (2) and (3) articulated above. I think that this is a clear case where in the human case our abstractions are constantly being swapped out in terms of better ones which are more useful (for example, compared to the ancient Greeks we have new representations for fundamental physics, not so much for dogs). I think it is very likely that once we had an AGI that crossed over the human threshold it would begin inventing novel mathematical structures and spinning up exotic physics unless for some reason its architecture can't do that, or we have somehow prevented this from happening. If interpretability tools are not equipped for this, it will probably come as a bit of a shock and break everything.

Specialized Concepts

Specialized concepts (you might also say "domain specific") is a category I am tentatively naming because I don't know of an official name. I pointing at the phenomenon where there are a lot of concepts and representations that we invent ad-hoc for particular purposes and we don't take them to be representing the "structure of reality" in the same way we do for physical concepts, nor do we take them to be pointing to something like "objective" natural kinds. This makes it much less clear that there is a pressure or convergence for a wide variety of cognitive architectures to learn them.

For example, when we invent the game of chess, we invent a bunch of novel representations for the different pieces and move states which transcend the particular instantiation of the pieces. When you invent the game of Chess and dub the piece of wood you are using as "pawn" you are not intending that to be name for that particular piece of wood, but minting a new concept which articulates the rules you are committing yourself to in treating any piece of wood as a pawn. With all of those concepts you now have a way of thinking about reality in a new way, rather than looking at a slab of wood you are "seeing" a game board and can think in terms of the movement of pieces. This is saving you a lot of computing power by abstracting away the features that aren't relevant for the game. But it would never occur to you to invent this kind of representation unless you were already in the mode of inventing abstractions for the game, the purpose of the game precedes the usefulness of the abstraction.

We are always inventing concepts like this on the fly for the various things we are doing. The reason I include this is that it seems likely to me that an AGI will begin spontaneously generating all sorts of novel conceptual structures for domains that we haven't conceptualized yet. This is another reason to think that interpretability might break down as the system grows more advanced.

Unfortunately both theoretical concepts and specialized concepts seem like they make up a decent portion of the conceptual space of cognitive systems. If some of the considerations I have suggested here materialize in full AGI this could pose a big challenge.

Suppose that an AGI does employ alien concepts, what does this mean for interpretability?

Useful interpretability already seems hard, and so anything that makes it more difficult brings down the probability that we will be able to do it successfully. If alien concepts are likely as we enter the region of the AI space where human abilities are outmatched, then if we want to continue interpreting the system we will have had to figure out some kind of automated way of doing this. In other words when it comes to understanding how an AI is making decisions or reasoning about the world we can either use transparency tools to understand how knowledge is represented in the system, and attempt to trace out how it is using that knowledge to make decisions. Or we could rely on automated ways of doing this like ELK or debate. Alien concepts might not matter much in the latter case (although it certainly makes the game harder), although I think it likely breaks the former one.