Phenomenological AI Alignment

1. Towards an Axiological Approach to AI Alignment
2. The World as Phenomena
3. Methods of Phenomenology
4. Form and Feedback in Phenomenology
5. Introduction to Noematology
6. AI Alignment and Phenomenal Consciousness
7. Computational Complexity of P-Zombies

Towards an Axiological Approach to AI Alignment

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.
This is a linkpost for https://mapandterritory.org/towards-an-axiological-approach-to-ai-alignment-4993d044d1b8

NB: Originally posted on Map and Territory on Medium, so some of the internal series links go there.

AI alignment research currently operates primarily within the framework of decision theory and looks for ways to align or constrain utility functions such that they avoid "bad" outcomes and favor "good" ones. I think this is a reasonable approach because, as I will explain, it is a special case of the general problem of axiological alignment, and we should be able to solve the special case of aligning rational agents if we hope to solve axiological alignment in general. That said, for computational complexity reasons having a solution to aligning rational agents may not be sufficient to solve the AI alignment problem, so here I will lay out my thinking on how we might approach AI alignment more generally as a question of axiological alignment.

First, credit where credit is due. Much of this line of thinking was inspired first by Paul Christiano's writing and later by this pseudonymous attempt to "solve" AI alignment. This led me to ask "how might we align humans?" and then "how might we align feedback processes in general?". I don't have solid answers for either of these questions—although I think we have some good ideas of research directions—but asking these questions has encouraged me to think about and develop a broader foundation for alignment problems that is also philosophically acceptable to me. This is a first attempt to explain those ideas.

Philosophical Acceptability

When I say I want a "philosophically acceptable" foundation for alignment problems, what I mean is I want to be able to approach alignment starting from a complete phenomenological reduction. Such a desire may not be necessary to make progress in AI alignment, but I would probably not be considering AI alignment if it were not of incidental interest to addressing the deeper issues it rests upon, so it seems reasonable for me to explore AI alignment in this way insofar as I also consider existential risk from AI important. Other people likely have other motivations and so they may not find the full depth of my approach necessary or worth pursuing. Caveat lector.

The phenomenological reduction is a method of deconstructing ontology to get as near phenomena as possible, then reconstructing ontology in terms of phenomena. The method consists of two motions: first ontology is bracketed via epoche ("suspension" of interpretation) so that only phenomena themselves remain, then ontology is reduced (in the sense of "brought back") from phenomena. The phenomenological reduction is not the same thing as scientific reduction, though, as the former seeks a total suspension of ontology while the latter accepts many phenomena as ontologically basic. In phenomenological reduction there is only one ontological construct that must be accepted, and only because its existence is

inescapably implied by the very act of experiencing phenomena as phenomena—the intentional nature of experience.

This is to say that experience is always in tension between the subject and object of experience. As such experience exists only in relation to a subject and an object, an object exists for (is known by) a subject only through experience, and only through experience does a subject make existent (know) an object. Put another way, phenomena are holons because they are wholes with inseparable parts where each part implies the whole. It is from suspension of ontology to see only this 3-tuple of {subject, experience, object} that we seek to build back ontology—now understood as the way the world is as viewed through phenomena and not to be confused with understanding the world as metaphysical noumena that may exist prior to experience—and thus all of philosophy and understanding.

Axiology

Axiology is traditionally the study of values (axias) and analogous to what epistemology is for facts and ontology is for things. I claim that axiology is actually more general and subsumes epistemology, ethics, traditional axiology, and even ontology because it is in general about combining experiences of experiences as object where we term such bracketed experiences "axias". Since this is not a frequently taken position outside Buddhist philosophy and there it is not framed in language acceptable to a technical, Western audience, I'll take a shortish diversion to explain.

Consider the phenomenon "I eat a sandwich" where I eat a sandwich. Bracketing this phenomenon we see that there is some "I" that "eats" some "sandwich". Each of these parts—the subject "I", the object "sandwich", and the experience of "eating"—is a sort of ontological fiction constructed over many phenomena: the "I" is made up of parts we might call "eyes", "mouth", and "body"; the "sandwich" is made up of parts we might call "bread", "mustard", and "avocado"; and "eating" is many experiences including "chewing", "swallowing", and "biting". And each of these parts is itself deconstructable into other phenomena, with our current, generally accepted model of physics bottoming out in the interactions of quarks, fields, and other fundamental particles. So "I eat a sandwich" includes within it many other phenomena such as "my mouth chews a bite of sandwich", "my eyes see a sandwich nearby", and even "some atoms in my teeth use the weak nuclear force to push against some atoms in the sandwich".

Now let us consider some experiences of the phenomenon "I eat a sandwich". Suppose you are standing nearby when I eat a sandwich, and you are engaged in the experience of watching me eat the sandwich. Although English grammar encourages us to phrase this as "you watch me eat a sandwich", we can easily understand this to be pointing at the intentional relationship we might formally write as "you watch 'I eat a sandwich'" where "I eat a sandwich" is a phenomenon nested as object within the phenomenon of you watching. We can similarly consider phenomena like "you read 'I write "I eat a sandwich"'" to describe what happens when you read my written words "I eat a sandwich" and "I believe '"I am hungry" caused "I eat a sandwich"'" to give a possible etiology of my sandwich eating.

To talk about the general form of such phenomena that contain other phenomena, we might abstract away certain details in each phenomenon to find the general patterns they match. When we say "you watch 'I eat a sandwich'", we are talking about a

subject, “you”, having an experience, “watching”, of the objectified (bracketed) phenomenon “I eat a sandwich”. Calling this bracketed phenomenon an axia, we can say this experience takes the form {subject, experience, axia}. Similarly “you read ‘I write “I eat a sandwich”’” is of the form {subject, experience, axia} although here we can deconstruct the axia into the nested phenomenon {subject, experience, {subject, experience, object}}, and “I believe ‘“I am hungry” caused “I eat a sandwich”’” is of the form {subject, experience, {axia, experience, axia}}. Even “I eat a sandwich”, which is really more like “I eat ‘I experience stuff-as-sandwich’”, has the form {subject, experience, axia}, and since the “I” we place in the subject is itself the phenomenon “I experience myself”, we can generally say that all experiences that we as conscious beings recognize as experiences are of one axia experiencing another.

This implies that axiology may be deeply connected to consciousness, and while I don’t wish to fully address consciousness here, we must say a few things about it that are relevant to axiology.

Recall that phenomenological reduction forces us to make only the ontological assumption that stuff experiences stuff. Stuff experiences other stuff differentially, so stuff appears to clump into clusters of stuff that have more, stronger experiences between the stuff in the cluster than the stuff outside the cluster. We call such clusters things, though note that all things have fuzzy boundaries that depend on how we choose to delineate clusters, and things are ontological constructions to help us reason about stuff and its experiences and not necessarily noumena. We may then say that things experience things to talk about the experiences connecting the stuff inside one cluster with the stuff inside another, and we call such stuff-to-stuff, thing-to-thing experiences “direct experiences”.

Now consider a thing we’ll label T. If this thing experiences itself as object—viz. there are “feedback” phenomena of the form {T, experience, T}—we say that T is cybernetic because it feeds experiences of itself back into itself. Since things are clusters of stuff experiencing other stuff within a cluster, all things are necessarily cybernetic, but we can distinguish among cybernetic things based upon how much feedback phenomena we observe within them. Thus, for example, we may say that rocks are less cybernetic than trees are less cybernetic than humans even though they are all cybernetic.

A cybernetic thing may experience itself experiencing itself. That is, there may be phenomena of the form {T, experience, {T, experience, T}} that occur when the stuff of T interacts with the stuff of itself as it interacts with the stuff of itself. We term such phenomena “qualia” and say that things experiencing themselves experiencing themselves are phenomenologically conscious. As with the definition of cybernetic, this implies that everything is conscious, although some things are more conscious than others, and most of the distinctions we care about are ones of how qualia are structured.

Since axias are the objects of qualia, they are the values or priors used by conscious things to engage in the experiences we variously call thinking, reasoning, and generally combining or synthesizing information. This is important because, although feedback and direct experience also combine information, conscious things *as conscious things* only combine information via qualia, thus conscious thought is entirely a matter of axiology. In this way axiology is to consciousness what cybernetics is to feedback and what physics is to direct experience because axiology is at heart the study of qualia. So if we are to say anything about the alignment of artificially conscious things, it will naturally be a matter of axiology.

Axiological Alignment

Note: In this section I give some mathematical definitions. These are currently prospective and may be revised or added to based on feedback.

Informally stated, the problem of AI alignment (also called AI safety and AI control) is to produce AIs that are not bad for humans. This sounds nice, but we need to be more precise about what "not bad for humans" means. I can find no standard formalizations of the AI alignment problem, but we have a few partial attempts:

- Stuart Armstrong has talked about the AI control problem in terms of an agent learning a policy π that is compatible with (produces the same outcomes as) a planning algorithm p run against a human reward function R.
- Paul Christiano has talked in terms of benign AI that is not "optimized for preferences that are incompatible with any combination of its stakeholders' preferences, i.e. such that over the long run using resources in accordance with the optimization's implicit preferences is not Pareto efficient for the stakeholders."
- MIRI researches have formally described corrigibility, a subproblem in AI alignment, in terms of utility functions.
- Nate Soares of MIRI has also given a semi-formal description of AI alignment as the value learning problem.

Each of these depends, in one way or another, on utility functions. Normally an agent's utility function U is defined as a function $U:X\rightarrow\mathbb{R}$ from some mutually exclusive options X to the real numbers. This, however, leaves the domain (pre-image) of the function ambiguous and only considers images (codomains) that are totally ordered. I believe these are shortcomings that have prevented a complete and rigorous formalization of the problem of AI alignment, let alone one I find philosophically acceptable.

First consider the domain. In toy examples the domain is usually some finite set of options, like {defect, cooperate} in the Prisoner's Dilemma or {one box, two box} in Newcomb-like problems, but in more general cases the domain might be the set of preferences an agent holds, the set of actions an agent might take, or the set of outcomes an agent can get. Of course preferences are not actions are not outcomes unless we convert them to the same type, as in making an action a preference by having a preference for an action, making a preference an action by taking the action of holding a preference, making an outcome an action by acting to effect an outcome, or making an outcome an action by getting the outcome caused by some action. If we could make preferences, actions, outcomes, and other things of the same type, though, we would have no such difficulty and could be clear about what the domain of our utility function is. Since for our purposes we are only interested in phenomenologically conscious agents, we may construct the domain in terms of the axias of the agent.

To make clear how this works let's take the Prisoner's Dilemma as an example. Let X and Y be the "prisoner" agents who can each choose to defect or cooperate. From X's perspective two possible experiences must be chosen between, {X, experience, {X, defect, Y}} or {X, experience, {X, cooperate, Y}}, and these qualia yield the axias {X, defect, Y} and {X, cooperate, Y}. Y similarly chooses between {Y, experience, {Y, defect, X}} and {Y, experience, {Y, cooperate, X}} with axias {Y, defect, X} and {Y, cooperate, X}. Once X and Y make their choices and their actions are revealed, they each end up experiencing one of four axias as world states:

- {X, defect, Y} and {Y, defect, X}
- {X, defect, Y} and {Y, cooperate, X}
- {X, cooperate, Y} and {Y, defect, X}
- {X, cooperate, Y} and {Y, cooperate, X}

Prior to defecting or cooperating X and Y can consider these world states as hypotheticals, like {X, imagine, {X, experience, {X, defect, Y} and {Y, defect, X}}}, to inform their choices, and thus can construct a utility function with these axias as the domain. From here it is straightforward to see how we can expand the domain to all of an agent's axias when constructing their complete utility function.

Now consider the requirement that the image of the utility function be totally ordered. Unfortunately this excludes the possibility of expressing incomparable preferences and intransitive preferences. Such preferences are irrational, but humans are irrational and AIs can only approximate rationality due to computational constraints, so any complete theory of alignment must accommodate non-rational agents. This unfortunately means we cannot even require the image be preordered and can at most demand that all agents can **approximately order** the image, which is to say they apply to the image an order relation, $\preceq$, defined for a set S such that for all $a \in S$ $a \preceq a$ and there exist $a,b \in S$ where $a \preceq b$. Humans are known to exhibit approximate partial ordering, where $\preceq$ is transitive and anti-symmetric for some subsets of the image, and humans and AIs may be capable of approximate total ordering where they totally order some subsets of the image, but lacking a proof that, for example, an agent orders its axias almost everywhere, we can only be certain that agents approximately order their axias.

Thus a more general construct than the utility function would approximately order an agent's axias. Let $\mathscr{A}$ be the set of all axias. Define an **axiology** to be a 2-tuple {A,Q} where $A \subseteq \mathscr{A}$ is a set of axias and $Q: A \times A \to \mathscr{A}$ is a **qualia relation** that combines axias to produce other axias. We can then select a **choice function** C to be a qualia relation on A such that {A,C} forms an axiology where C offers an approximate order on $\mathscr{A}$.

Given these constructs, we can attempt to formally state the alignment problem. Given agents X and Y, let {A,C} be the axiology of X and {A,C'} the axiology of Y. We can then say that X is **axiologically aligned** with Y—that {A,C} is aligned with {A,C'}—if for all $a,b \in A$, $C'(a) \preceq C'(b)$ implies $C(a) \preceq C(b)$. In English what this says is that one axiology is aligned with another if the former always values the same axias as much as the latter.

An immediate problem arises with this formulation, though, since it requires the set of axias A in each axiology to be the same. We could let X have an axiology {A,C} and Y an axiology {A',C'} and then define axiological alignment in terms of $A \cap A'$, but aside from the obvious inadequacy to the purpose of alignment if we exclude the symmetric difference of A and A', the axias of each agent are subjective so $A \cap A' = \varnothing$. Thus this definition of alignment makes all agents vacuously aligned.

So if axiological alignment cannot be defined directly in terms of the axias of each agent, maybe it can be defined in terms of axias of one agent modeling the other. Suppose X models Y though qualia of the form {X, experience, {Y, experience, axia}}, or more properly {X, experience, {{X, experience, Y}, experience, axia}}}. Let these axias {{X, experience, Y}, experience, axia}} form a subset A_Y of A. Within A_Y will be an axia {{X, experience, Y}, experience, C'}, and we identify this as C'_X. Then we might say X is **weakly axiologically aligned** with Y if for all $a,b \in A\_Y$, $C'\_X(a) \preceq C'\_X(b)$ implies $C(a) \preceq C(b)$.

As hinted at by “weak”, this is still insufficient because X may trivially align itself with Y by having a very low-fidelity model of Y, so Y will want some way to make sure X is aligned with it in a way that is meaningful to Y. Y will want X to be corrigible at the least, but more generally will want to be able to model X and assess that X is aligning itself with Y in a way that is satisfactory to Y. So given that X is weakly aligned with Y, we say X is **strongly axiologically aligned** with Y if for A’_X, the set of axias {{Y, experience, X}, experience axia}, and C_Y, the axia {{Y, experience, X}, experience, C}, for all a,b∈A_X, C(a)≤C(b) implies C_Y(a)≤C_Y(b).

This is nearly a statement of what we mean by the AI alignment problem, but the choice of a single agent Y seemingly limits us to aligning an AI to only one human. Recall, however, that everything is at least somewhat phenomenologically consciousness, so we can let our agent Y be a collection of humans, an organization, or all of humanity united by some axiology. Traditional axiology leaves the problem of combining individuals’ axiologies to ethics, so here we will need a sort of generalized ethics of the kind sketched by coherent extrapolated volition, but for our present purposes it is enough to note that axiology still applies to a meta-agent like humanity by considering the qualia of humanity as a phenomenologically conscious thing. Thus, in our technical language,

> The AI alignment problem is to construct an AI that is strongly axiologically aligned with humanity.

Research Directions

Stating the problem of AI alignment precisely does not tell us a lot about how to do it, but it does make clear what needs to be achieved. In particular, to achieve strong alignment we must

- construct an AI that can learn human axias and make them part of its own axiology,
- understand how to assess with high confidence if an AI is aligned with human axias,
- and develop a deep understanding of axiology to verify the theoretical correctness of our approaches to these two tasks.

Current alignment research focuses mainly on the first two issues of how to construct AIs that learn human axias and do so in ways that we can verify. The axiological alignment approach does not seem necessary to continue making progress in those areas now since progress is already being made without it, but it does provide a more general theoretical model than is currently being used and so may prove valuable when alignment research advances from rational agents to agents with bounded or otherwise approximate rationality. Further, it gives us tools to verify the correctness of AI alignment work using a rigorous theoretical framework that we can communicate clearly within rather than hoping we each understand what is meant by “alignment” in terms of the less general constructs of decision theory. For this reason alone axiological alignment seems worth considering.

I anticipate that future research on axiological alignment will primarily focus on deepening our understanding of the theory laid out here, exploring its implications, and working on ways to verify alignment. I also suspect it will allow us to make progress on specifying humanity’s choice function by giving us a framework within with to build it. I encourage collaboration and feedback, and look forward to discussing these ideas with others.

The World as Phenomena

This is a linkpost for https://mapandterritory.org/the-world-as-phenomena-47d27d593016

NB: Originally posted to Map and Territory on Medium, so some of the internal series links go there.

NB: I made some mistakes in this post about my use of academic philosophy terms in ways that I would no longer endorse. I've let it stand for historical purposes, but of particular import is to note that "existential phenomenology" fails to capture all of my perspective well, and my understanding of idealism in this post is rather poor, since my perspective is well described, in part, by idealism, specifically phenomenalism.

My introduction to axiological alignment is just that—an introduction. Out of necessity it covers the philosophical foundations of my thinking on AI alignment, but it doesn't explore those foundations in much detail. I could just keep writing about axiological (now noematological) alignment and AI, but I suspect I would lose everyone, so first I'm going to go back and explicate the background theory as thoroughly as seems necessary. We'll begin at the beginning, employ examples and metaphors, and explore, in order, the foundations of phenomenology, phenomenological methods, the phenomenological reduction, feedback, qualia, and finally noematology.

I think of phenomenology as the philosophy of the adept beginner. By this I mean it is the philosophy of the person skilled in all the techniques of philosophy from logic to observation to intuition who nevertheless chooses to approach philosophy as if they were a beginner so that, rather than trying to explain anything, they rest on the questions until they answer themselves. Phenomenologists are not the first to try this—Socrates, Zhuangzi, Nagarjuna, Descartes, and Hegel are among our predecessors—but we have the advantage of standing on their shoulders. Let's see how far we can see!

Consider the fundamental question of philosophy, "why?".

> "Okay, why?"
>
> *"Why" what?*
>
> "Why is anything?"
>
> *What is this "anything" you refer to?*
>
> "The world. Why does the world exist?"
>
> *Does it? How do you know?*
>
> "I'm in it!"
>
> *Oh, so why are you in the world?*
>
> "I…I don't know. I just am."

It's from not knowing why "I am" that we start because if we follow any line of questioning far enough we'll eventually need to know what this "I" is to give a

complete answer. Even if any particular answer doesn't seem to depend on "I", answers are given and understood by an "I", and we can always ask of any answer "Why do I think that?", so we have to address "I" sooner or later. Of course, we also have to deal with the whole world sooner or later, so choosing to start with "I" is also a pragmatic choice because it's something we are each quite familiar with and have privileged knowledge of, and as we'll see later there is something epistemically irreducible about "I" that makes it interesting.

Now depending on your background and education, you might be tempted to dismiss this idea of starting with "I" and jump straight to taking an objective approach. I understand the appeal, but remember we are working from a beginner's perspective, and even the logic of science must begin with the subjective. If we don't actually need "I" and it evaporates as we develop our understanding, then so be it, but beginning as naifs we must include the naive conception of "I" for now and only remove it later if we can fully account for it.

Continuing our line of questioning,

> *How do you know you are? Don't think too hard; just give the obvious answer.*
>
> "I see and hear. I think. I remember. I experience."

If we take up the kind of the radical philosophical skepticism introduced by the Ajnana school and Pyrrho, one of the few things we can claim to know is that we experience. It may not be clear who we are, what we experience, or how we experience it, but to know is to put the "I" in relation to the world and we call that relationship "experience". Existential phenomenology is the philosophy that develops if you choose to assume experience is the only source of knowledge.

There are other options. Platonism, for example, chooses differently and permits the existence of direct knowledge without the need for it to be experienced. Some strains of Gnostic and Buddhist thought suppose there is only direct knowledge and experience is an illusion. And solipsism rejects nearly everything we think of as knowledge, granting only the existence of the "I". All of these are possible ways to address "why am I?", so why phenomenology?

Two reasons. First, modern physics makes it abundantly clear that there is probably no direct knowledge. We get no faster than light travel, no true spooky action at a distance, no nothing! We always pay for our lunch, even when it's free, and if there is direct knowledge, then we've gone a suspiciously far way to understanding the physics of our universe without discovering it. At the same time, what physics does show us is that information only exists when it's moving, which is to say when measurement happens, and thus appears tightly bound to the mechanisms of causality. So taken in whole, physics paints the picture of a world where the only way to know anything is through experience.

Second, parsimony. Reasonable people may disagree, but my take is that the phenomenological perspective—that experience is the only source of knowledge—is the least complex solution that is able to fully address the question of how we know of our own existence. If we try to get away with less complexity and say experience is not necessary, I think we fail to adequately explain why it appears to us that we have a shared, objective existence that extends beyond our own knowledge. And if we try to demand more complexity by, for example, assuming the existence of direct knowledge, I think we get a gear that doesn't turn the machine and so can leave it out of our understanding without losing anything. Thus, by granting experience sole

ownership of knowledge generation, we produce a “just enough” explanation that neither has anything missing nor anything extra.

I’ll have more to say on parsimony when we look closer at the phenomenological reduction, so for now let’s wrap up our dialogue.

> *So you experience, but what if you take away the “you”? Can there be unattached experience?*
>
> “…no? …no. No, there can’t.”

Experience is always experience of something, by something. That is, it is always that some subject experiences some object. In this way we say that experience is intentional because experience is directed and exists in tension between the subject and object the same way a rope may be held in tension between two posts. If we take away the subject or the object the experience falls away the same way the rope would go slack if we took away one of the posts, and if we take away the experience then there’s nothing connecting the subject and object so they would just be things, not subjects or objects, just as our two posts would be disconnected without the rope stretched between them. Taken together we call this relation of subject, experience, and object a phenomenon.

Suppose we take a specific phenomenon, like “dog barks at car”, and try to show it is not intentional by separating out its parts. Let’s start with the subject. If we take out the dog we have only “barks at car”, but even without the dog we must imagine something is doing the barking. This gives us with a pattern that can be matched by a subject, but then we are still describing the experience of something barking at a car, so we still have a complete phenomenon, albeit one with a thought as the subject rather than a thing. As we’ll see when we address qualia, the subject being a thought poses no problem because the subject, when experienced as the subject of a phenomenon, is always a thought anyway.

If we can’t fully remove the subject, what about the object? Taking out the car nominally gives us “dog barks” and, at least in English, this is a valid construction where the experience/verb has no object. But is there really no object? Consider what barking is: the dog’s bark happens when the dog uses its body to create compression waves in the air that we identify as barking, so “barks” has an implicit object like the air or the world into which the dog does its barking. Thus we can’t really remove the object either, although we can change our frame of reference to see the object differently.

Since we can’t remove the subject and we can’t remove the object, we obviously can’t remove both and get a pure experience of barking, but what if we remove the experience and attempt to hold the subject and object as things-in-themselves? This almost seems possible, but ask yourself how you know about the dog and the car. Since we assumed that experience is the only source of knowledge, it must be that something experiences the dog and car to know about them, and in this case that thing is you! Maybe the dog and car can exist without being known, but in existential phenomenology their existence is fundamentally unknowable if they are not

experienced by a subject. So even if the dog is not related to the car by the dog's barking, the dog and the car are objects of your own experience and thus part of some phenomena. If they were not they would be literally unknown.

Thus we see that phenomena, although they have parts, are not divisible, and all knowledge exists as phenomena. This is the keystone of the phenomenological perspective.

I see. You were secretly a phenomenologist all along!

Don't be too surprised if you find yourself nodding along and thinking this all sounds obvious. It is! The trouble comes with the philosophical bullets you'll be asked to bite when taking this radically naive view.

Consider the question of whether a thing exists in its own right independent of phenomena. That is, do things exist if they are not experienced? For example, suppose there is a star so far away that the light from it will never reach us. Since it is outside our light cone we will never experience evidence of it, so in what sense can this hypothetical star be said to exist? From the phenomenological perspective, we might say that the *idea* of such a star exists but that the star itself does not since we can't interact with it.

Or consider particle physics. In what sense are quarks or strings or anything else "real"? Scientific realism holds that physical models, to the extent they are accurate, describe things as they really are—viz. if atomic theory correctly describes the world then the world really is made up of actual atoms. But this is to suppose a direct knowledge of the world which we can uncover through scientific inquiry. A phenomenologist might instead say we can investigate the ontic being of a thing, but only because we have knowledge of the ontological and through that may infer something of the metaphysical. This leaves phenomenology compatible with physicalism, but finds it decidedly opposed to realism and, in its existential form, idealism.

If none of that sounded too weird, when we get around to discussing qualia and consciousness we'll find that existential phenomenology implies something like functionalism and definitely implies panpsychism. I won't try to convince you of that now, but know that our humble choice to accept that all knowledge comes from experience will lead us to places far off the beaten path. It should be an exciting ride!

Methods of Phenomenology

This is a linkpost for https://mapandterritory.org/methods-of-phenomenology-e2f936651ff

NB: Originally posted on Map and Territory on Medium, so some of the internal series links go there.

In the previous post we looked at phenomenology and the thinking that motivates it. We saw that it is based on taking a naive, skeptical, beginner's view to asking "why?" and choosing to address the question using only knowledge we can obtain from experience. We also saw that experience is intentional: that it is directed from subject to object and forms an inseparable, arity 3 relation called a phenomenon. Taken together this gave us a feel for the shape of phenomenological philosophy and allowed us to glimpse some of the consequences of taking this view seriously.

We now have the context to begin exploring phenomenology's details, and the first detail to explore is the methods of exploration themselves because phenomenology is highly integrative and our assumptions—that we know the world only through experience and that experience is intentional—determine what sorts of methods we can use. Thus if we are to approach phenomenology we must first gain some familiarity with its practices.

Now if I'm honest there is only really one method of phenomenology—the phenomenological reduction—but that's a bit like telling you that the only method of decision making is Bayes's Theorem: in a sense it's true, but it's not likely to help you understand anything in the same way that telling you that you are already enlightened doesn't make you enlightened. To explain how phenomenologists think, it's more useful to talk about a broad base of specific methods and through them approach the reduction proper. Luckily, there are many methods, and you are almost certainly already familiar with several of them, so let's take a look!

Science

It might seem a little surprising to you that science is a method of phenomenology since, historically, phenomenology emerged in part from Husserl noticing the inadequacy of natural science for addressing conscious experience, and, in the latter-half of the 20th century, phenomenologically inspired thinkers in the post-modern movement sometimes took anti-scientific stances. But Husserl viewed science as part of phenomenology, and some consider phenomenology a science of consciousness, so there's more to the relationship between science and phenomenology than some surface level antagonism.

Like phenomenology, science starts from a place of empiricism—the idea that knowledge is obtained through experience and observation. To the extent that science is systematized empiricism, phenomenology is a science, but "science" usually refers to a more specific practice of the scientific method with certain standards of evidence that phenomenology doesn't always hold itself to. In particular, science considers only phenomena where subjects and intentionality can be ignored because the objects and the experiences of them remain relatively unchanged between subjects. We call such truncated, replicable experiences of objects objective or natural phenomena.

One of the foundational issues of science is to decide exactly what the criteria for objective phenomenon are, but generally objective phenomena are those that describe sufficiently similar experiences of objects no matter who or what the subject is, like the way a beam of light will be experienced as having the same wavelength no matter who sees it or what measures it. By limiting itself to these objective phenomena, science is able to make predictions about the world that it expects to hold for all subjects, which is to say that what is true of a few phenomena will be true of all similar phenomena. This let's us uncover patterns science calls theories that make strong predictions about the world.

As so far described, science is compatible with phenomenology and allows us to make much more confident statements about the world when objective phenomena are available than when they are not. But objective phenomena cannot always be reliably constructed. Because objective phenomena are not actually phenomena but patterns of statistical regularity observed over many phenomena, there is necessarily information lost in the creation of objective phenomena, limiting what can be known through them. This might seem like an obscure, technical issue for philosophers of science, but it has implications for when science, in the sense of understanding the world through the use of objective phenomena, is an appropriate method.

Phenomenology views science as extremely useful but sees it running aground the closer it gets to exploring topics where the intentional nature of experience matters and objective phenomena are less available, such as in the study of consciousness. Thus science is a great method for exploring questions of physics, chemistry, and biology; pretty good for studying economics and archeology; workable in psychology and anthropology; and of limited direct usefulness in philosophy and philology. For those topics where science cannot cover all the epistemology ground, other phenomenological methods are necessary.

An Aside on Scientism, Irrationality, and Their Kin

Now I'd rather not have to write this part but I suspect some readers may be upset at me for presenting the relationship between phenomenology and science as prosaic. The technical issue of determining how much we can figure out with science alone has and does get mixed up with all sorts of discussions about other things, so I think it's worth saying a few words about this to at least acknowledge the issue and direct you to additional reading if this topic is of interest.

Humans are political animals, so when there is disagreement on something it often sparks or gets sucked into a larger battle between groups. One of these battles is along a dimension we might call "rationality" between those who value the modern worldview and those who don't. The details get complicated, but you can basically imagine it as if there were two political parties vying for control of a country, the Pro-Rationality Party and the Anti-Rationality Party, and it's into this milieu that phenomenology and science are thrown.

The Pros claim science for their own, so the Antis reject it. Phenomenology says science is useful for understanding many things but not literally all things, so the extremists on the Pro side reject phenomenology for being "impure". The Anti side then takes phenomenology in and plays up the limits-of-science thing while downplaying the usefulness-of-science part. As a result phenomenologists more often find themselves having to defend their ideas against material realism, scientism, and other ideas on the Pro side and less against irrationality, mysticism, and other incompatible positions on the Anti side. This creates a skewed picture that implies

phenomenology is anti-science by association, and it doesn't help that some phenomenologists, being humans, may actually take up sides in this debate.

But ultimately the Pro/Anti battle is more about how humans relate to ideas than the ideas themselves, and methods like debate magnify this confusion. Thankfully, phenomenologists and other philosophers have an alternative to debate that functions better at collaborative truth seeking: the dialectic.

Dialectic

Philosophers have a special way of talking to each other in good faith that cuts to the heart of their disagreements. Once they find these disagreements they can build toward mutual understanding and possible agreement. We generically call this process dialect and I've written about it before:

Debate and Dialectic
*Election season is over in the US, but folks are still talking about how divided political conversation is. We hear…*mapandterritory.org

In case you don't want to click the link, the short version is that the dialectical method is to consider a position or idea, called the thesis, find something that contradicts it, an antithesis, and then try to find a synthesis of thesis and antithesis that sublimates/overcomes them with new understanding. That new understanding becomes a new thesis, and the process repeats until it converges to consensus or diverges to logical inconsistency.

Dialectic differs from debate so long as the antithesis and synthesis are formed in good faith. With good faith, dialectic can get at the heart of a dispute to find agreement, but with bad faith dialectic devolves into debate and drives a wedge between people and ideas. Since phenomenology gives primacy to phenomena, including those non-objective experiences we might call "subjective", it strongly encourages good faith and is able to make extensive use of dialectic as a tool for building understanding.

There is not always disagreement or apparent contradiction to power a dialectic, though. In those cases phenomenologists must explore the world in other ways, and more closely examining the phenomena themselves often yields dividends.

Hermeneutics

When we write, speak, or otherwise communicate, we engage in an act of creating phenomena for others by giving them objects to experience. We can try to anticipate how they will experience these objects—to predict the phenomena our audiences will find themselves subject to when they read our writing, hear our words, or see our art—but there will inevitably be variation in their experiences. This means that for all experiences of the same object there will be different experiences had by different subjects. This opens up the opportunity to compare and study the differences in experiences, and we call this study hermeneutics.

Although technically it is possible to perform hermeneutics on phenomena of non-conscious subjects, we generally consider that practice a part of applied science or engineering, so hermeneutics generally refers to the process of interpreting the experiences of conscious subjects. Originally hermeneutics primarily focused on

interpretation of sacred experiences, especially of messages believed to have been sent by the gods, but Heidegger generalized the notion within phenomenology to interpretation of experience and developed the hermeneutic circle as his primary philosophical technique. Philologists then mixed Heidegger's philosophical hermeneutics with their own methods and developed techniques we now think of as literary criticism, historical analysis, and other methods of critical study in the humanities.

I think of hermeneutics as a kind of meditation on the experiences of others where people report their experiences and we think on those reports to create our own experiences of them. We only have access to our own experiences, but from our experiences we can reason about the world that made possible the experiences of others and so gain partial, indirect knowledge of objects of experience we never experienced ourselves. In this sense hermeneutics is what we do whenever we read a book, listen to a friend talk, or empathize with the experiences of others.

We can similarly think of meditation as hermeneutical analysis of our own experiences, but this would be selling meditation short because, unlike when analyzing the reported experiences of others, we are the subjects of our own experiences and can, at least in theory, know more about them. Turning this theory to practice is not easy, though, so meditation is a method of phenomenological epistemology worth exploring on its own.

Meditation

"Meditation" is a word with a lot of meanings. In one sense it means focused thinking on a topic, and you might say my writings are often meditations of this sort. There's another sense in which meditation is the practice of entering trances and other altered states of consciousness, possibly associated with spiritual experiences, and while this is interesting because it may produce qualia not otherwise generated, it's not a phenomenological technique so much as a source of capta. Instead the sense in which we care about meditation is as a method of cultivating awareness of the world and our interactions within it so that we may learn everything we can from our experiences.

There are many specific meditative practices that can serve phenomenological purposes. For example, the meditation of early phenomenologists was heavily influenced by yogacara, I practice zazen, and any technique that teaches the ability to observe phenomena without interpretation will work. The key is learning to withhold judgement so that, as much as possible, the world may be seen as it is. From gaining such a clear picture of the world we may start our naive, skeptical, beginner's investigation of it.

Being skeptical, it's fair to ask how much value we can derive from meditation. After all, psychology is littered with disproved theories that drew much of their evidence from introspection, so there seems reason to be suspect of anyone claiming knowledge solely based on their own experiences. But just as science abandoned those theories when their evidence did not reproduce, the phenomenological framework similarly does not ask you to accept the evidence provided by others (or yourself!) blindly. If someone reports an experience that seems false to you in some way, you should try to understand it, and if you desire to know more you should try meditating on similar experiences yourself to see what you learn. If you get different results than others, that can be a starting point for dialectic and hermeneutics.

Thus it's important to be clear that meditation is like science, dialectic, and hermeneutics in that it does not stand on its own. Meditation cannot give us perfect knowledge even as it helps us to approach the limits of our knowledge imposed by the intentional nature of experience. But how close can we get to those limits? Husserl believed it was possible to get so close as to feel yourself transcending them, but any such feeling of transcendence must itself be an experience that can be suspended and examined, so it seems at best we can reach an equilibrium of continuously experiencing the experience of experiencing experience. To make sense of such deeply self-referential phenomena, Husserl developed the phenomenological reduction, the foundational method of phenomenology.

The Phenomenological Reduction

All phenomenological methods are expressions of the phenomenological reduction. They're not like this because they were designed this way: most phenomenological methods predate the idea of phenomenology itself. Instead, the phenomenological reduction is the core movement available to us as we explore the world via phenomena, and so all other methods are naturally expressions of it. That we do not always use the naked reduction directly reflects the difficulty of carrying out the reduction in full.

The reduction is not very easy to describe, either. It consists of a single movement with two motions—epoche and epistrophe. "Epoche" is the Greek word Husserl used to refer to the process of suspending, stepping back from, or bracketing an experience so that it may be examined, and epistrophe is the dual or reverse process of epoche where we return, reintegrate, or reduce our understanding back from suspension. Confusingly Husserl didn't use the term "epistrophe" to match "epoche" but instead referred to epistrophe as "the reduction proper" (German: *das eigentliche Reduzieren*, "the reduction in its own light") based on the original Latin meaning of *reducere* from *re-* meaning "back" or "again" and *ducere* meaning "lead" or "bring". Given the confusion this invites both because it gives too similar names to the method and one of its motions and because "reduction" is now philosophically cognate with reductionism, I choose to use "epistrophe" instead.

Notice that I called epoche and epistrophe motions and not steps or parts. This is intentional because the reduction is a complete movement where one motion naturally follows the other. You might think of epoche as breathing in, epistrophe as breathing out, and reduction as breathing: you have to breathe in and breathe out to breathe, if you breathe in you necessarily breathe out, and if you breathe out you will almost certainly breath in again. Thus although we may talk about the two motions separately, they fundamentally imply one another.

To see the reduction at work, let's perform its motions on a classic example from phenomenology, seeing a cup.

We first perform epoche by suspending the action of seeing the cup to experience the phenomenon of seeing a cup as phenomenon. That is, we quote the phenomenon “I see a cup” so that we can consider it apart from our participation in it. From there we might bracket the phenomenon further to find that, for example, what we think of as a cup is actually our mind interpreting particular sensations as a cup, and those sensations are themselves further phenomena of cells in our body producing chemical-electrical signals in response to light. We can continue epoche until the phenomena we wish to examine have been bracketed or we lack the insight to see a further suspension.

We then move into epistrophe to reconstruct what we deconstructed via epoche having gained a broader perspective. Before a cup was just a cup; now we can see a cup as a construction of multiple layers of phenomena adding bits of meaning—what we might also call ontology, categorization, modeling, pattern matching, or the map—leading up to an experience of seeing a cup as a cup. Along the way we get a picture of how a cup comes to be and how it is differentiated from other things, and if we still do not see reality as clearly as we need to after this, we move back into epoche to begin the reduction again.

The cycle of epoche and epistrophe forces us to be parsimonious. If we leave in epicycles or other sorts of complex assumptions during epoche or construct them during epistrophe, we will merely find ourselves needing to further bracket our perceptions until we can explain them to ourselves. And if we repeat the reduction long enough and take our thinking far enough, we end up doing what Husserl said made the reduction “radical”: we gain gnosis of consciousness and being. We’ll return to issues of consciousness soon enough, so for now let’s wrap up by seeing how reduction connects all of our methods.

Being trained in mathematical thinking, I often engage in a formal version of the phenomenological reduction to make my thinking clear (cf. how I showed my epoche in the introduction to phenomenological AI alignment). Husserl preferred to practice radical self-meditation, as he put it, to perform the phenomenological reduction. Heidegger’s hermeneutic circle is a thin layer on top of phenomenological reduction, focused on seeing things alternately as made of parts and as wholes rather than as a movement between epoche and epistrophe. Dialectics often play out as epoche until they converge and enable epistrophe. And science most of all cherishes epoche as its method of suspending judgement to see the world as it is so that theories may be developed under formal rules of epistrophe. All, when done from a phenomenological perspective, embody the motions of the reduction and *the* method of phenomenology.

Next time we’ll begin exploring aspects of the world from a phenomenological perspective to prepare us for talking about noematological alignment in AIs. See you then!

Form and Feedback in Phenomenology

This is a linkpost for https://mapandterritory.org/form-and-feedback-in-phenomenology-d44f4e5c72b3

NB: Originally posted on Map and Territory on Medium, so some of the internal series links go there.

Now that we've covered phenomenology's foundations and its methods, we're almost ready to begin addressing AI alignment from a phenomenological perspective, but before we can start talking about AI or how to align it we need to build a clear understanding of what AI is and what aligning it would mean. To do this we need to understand noematology, which requires understanding consciousness and qualia, which requires understanding feedback, which requires understanding things and time. That's a lot to tackle at once, so we'll split the inferential chain and start where many philosophers have started—consciousness. Specifically, we'll begin with a phenomenological reduction of self experience.

To start our reduction, let's first bracket the experience of self as a phenomenon in a clear way that makes precise what we consider its subject, experience, and object. English grammar encourages us to phrase this phenomenon as "I experience myself", but in normalized form we instead denote it {I, experience, I} to mean that the subject, I, has some experience of itself. This immediately invites many questions. What is the I? How do we know it's both the subject and the object of this phenomenon? Does the I know it's experiencing itself or is it only experiencing itself without knowing what it's experiencing? We'll need to address all of them, but since they all ask something about the I, we'll begin by bracketing it and exploring its role as intentional subject and object.

Things

It's tempting to start by asking what kind of thing I is, but this immediately exposes an assumption we need to suspend—that I is a thing. Epoche of this idea is often a jumping off point for considering the essential nature of (or rather the lack of essence of) the self, and that's a topic we'll cover, but to address the I now would be to put the cart before the horse because doing so assumes we have a solid understanding of what a thing is! That may sound silly if you aren't used to philosophical discourse, but, as we'll see, understanding just what we mean when we talk about a thing is the cornerstone of complex phenomenological analysis.

It's a little hard to talk about what a thing is, though, because we naively think of things as ontologically basic. Look in a dictionary and you'll find most definitions of "thing" are circular: they define things as objects, items, entities, articles, or artifacts. The best I could find is the circumlocution "that which can be described", but this still leaves us with the trouble of how to specify what "that" is. Even the etymology of "thing", while hermeneutically interesting, doesn't much help us figure out what we mean when we talk about things. The epoche of "thing" seems to be our first real phenomenological challenge.

Rather than deal with things in the abstract, let's begin by reducing something more concrete. Most people would agree that a cup is a thing. Why? Well, they might say, a cup is distinct from its surroundings. For example, you can pick a cup up off a table and the cup remains a cup whether it's on the table on in your hand. But to say that the cup remains a cup is to suppose we already know what a cup is, so being a bit more skeptically cautious we step back and say only that we observe a cup on the table and continue to observe a cup when we pick it up and hold it in our hands. But even this is not cautious enough because we've still assumed that we can identify the cup, so we must also bracket the process of identification. Only then are we finally left with a question we can begin to answer: how do we see a cup from the pure sense data of experience?

One specific answer we might give is that our brains detect patterns based on the light that reflects off the cup and use those patterns to form a category. Then through a complex process of learning that we don't fully understand, we develop an association between those visual patterns and the sound of someone saying "cup", and then also the squiggles we use to write "cup", so that we come to have a grouping of patterns we assign the label "cup". This gives us a way of forming mental categories both for specific cups and cups in general, and by a similar process categories for other patterns we find in our experiences. Combined these categories create a logical structure we apply to the world as we observe it that we call ontology.

As with cups, so too with all things, thus we can say that things are ontological categories. But ontology does not seem to be a full accounting of thingness because it depends on the phenomenological subject. That is, ontology can only identify a thing via phenomena and gives the thing no existence independent of our experience of it, yet cups seem to continue to exist even when they are locked away in cupboards. Thus our naive sense of thingness seems to extend beyond ontological thingness, so it appears there must be some way in which things exist independent of conscious experiences of them.

Indeed, many of my philosophical forebearers agree. Kant called this sense in which things exist on their own the things-in-themselves or noumena as apart from phenomena, Heidegger referred to it as ontic being as opposed to ontological being, and Sartre distinguished between things and things as things. Each, I think, is looking for a way to talk about the natural organization of stuff that we observe as patterned prior to knowing that it's patterned, especially since it seems that our individual, independently constructed ontological categories are highly correlated with each other in a way that suggests there is an external reality that exists independent of our knowledge of it. In other words, they are trying to address the intersubjectivity of thingness.

My own take on the metaphysics of the ontic nature of things, for what it's worth, is as follows. The world is made up of stuff. That stuff might be particles, strings, fields, part of the universal wave function, or something else, but whatever it is this stuff interacts with other stuff and phenomenologically we call this interaction "experience". Further, stuff experiences other stuff differentially, so stuff metaphorically clumps into clusters of stuff that have more, stronger experiences between the stuff in the cluster than the stuff outside the cluster. I call such clusters ontic things, though note that we only know ontic things as ontological things through our experience of them, so we necessarily lack perfect knowledge of the ontic given our assumption that our only source of knowledge is experience. Nonetheless we can often be confident enough in our assessment of the ontic and ontological to act as if ontological categories point to ontic clusters of stuff and as if both describe the same naively conceived thing. Taken together the ontic and ontological give us a sense of the forms of the world even though it is properly formless.

Using this sense of what a thing is—an ontological category describing an ontic clustering of stuff in the world—we can begin epistrophe of our earlier epoche of I as thing. To say that the I is a thing is to say that we observe some pattern in the world that we identify as being the source of experience and give this thing the name "I". Thus when the I experiences itself, when {I, experience, I}, the I is experiencing itself as a thing it identifies as I. This addresses the first of our questions— "what is the I?"—but what about the second? How do we know that this same I is both the subject and the object of the phenomenon?

Time

Let's suppose, for the sake of argument, that when we talk about the phenomenon {I, experience, I} we intentionally mean to say by using the same lexicographical term, "I", for subject and object that the subject and object are exactly the same thing, viz. the I is experiencing exactly itself without any differences. This would imply that self experience cannot change the I since the I as subject cannot be modified by experiencing the I as object

else the I would not be experiencing exactly itself, but this seems at odds with the evidence from our own lives where self experience not only can change us but is often a leading force for change, so the subject and object must not be exactly the same. What is their relationship, then, that we are led to think of self experience as {I, experience, I}?

A first approximation might be to say that whatever it is that we identify as the I is not the entirety of I. That is, perhaps there is some “core” part of the I that persists across experiences associated with some “outer” part of the I that changes so that [I, experience, I] is really shorthand for [core-I, experience, core-I and outer-I] and this phenomenon produces a modified outer-I. Such a core-I is essentially a soul, and it does a better job of fitting the evidence that the self changes even as it seems to remain connected to its past than supposing that the entirety of the I never changes, but to me it seems insufficiently parsimonious to suppose that the I has two parts, especially since this theory would suggest that every thing that changes, not just the I, would have a “soul” of some kind.

We could instead take the I to consist only of the outer, mutable part and for the core, apparently immutable part to be a pattern within the outer-I that remains stable over time. Then when we say that {I, experience, I} what we mean is something more like {I’, experience, I} where I’ is I with a a specific experience of itself added in. I, I’, and their successors may experience their precursors many times forming a long chain of things stretching into the past that we associate with I, so writing {I, experience, I} is a handy way to elide the many phenomena and iterations of the I that we are actually considering. This theory has fewer moving parts than supposing the I has a soul and does not imply pansoulism, so it seems a decent way of accounting for the identification of the object with the subject in self experience.

Only, notice that above we had to use a concept of time, both explicitly and implicitly, to give our description. Even when we supposed there was a soul we needed time to address issues with the outer self. If the I were immutable this would not be an issue, but to talk of precursor and successor Is is to talk about the past and the future. So if we are to accept that the I can change, we must have an understanding of what that change is over or through. After all, to persist is literally to stand through, and if time is to be that thing through which the self stands, we’ll want to consider what time is.

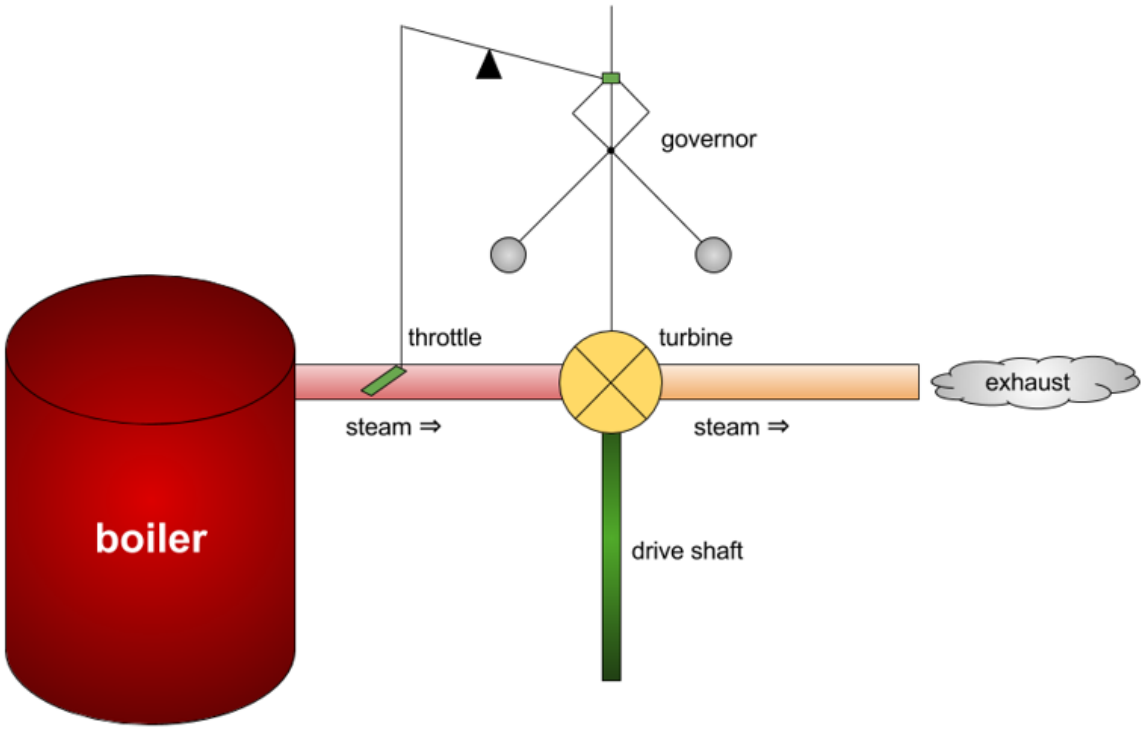
Husserl and Heidegger both gave considerable thought to the experience of time and found it to be not a thing but an aspect of intentionality. That is, time is not a dimension as it was often conceived of in Newtonian physics, but rather a consequence of the intentional nature of experience. That we can think of time as an objective medium or external reference frame through which we move is an ontological convenience we construct to make it easier to consider time as a phenomenological object, but in fact time is epiphenomenal and it's only through phenomena that we come to experience the world as if it had time. This makes phenomenological theories of time examples of causal theories of time where causality creates a pattern we call time rather than metaphysically basic time creating causality, which is nice because modern physics seems to agree that we need a causal theory of time.

Understanding time as an effect rather than a cause has dramatic ramifications for our understanding of things. To talk as if a thing existed in the past or will exist in the future is really to say that phenomena contain patterns identified as a thing and that this pattern suggests it was also present in past phenomena. For that matter, the notions of "past phenomena" and "future phenomena" are themselves patterns in present experience that suggest the existence of phenomena that created the world as it is presently experienced and suggest the existence of phenomena that will be created based on the world as it now seems. Consequently I tend to think of time as an ordering relation that forms a poset over the set of phenomena, but we need not commit ourselves to any particular interpretation so long as we understand that things lack permanence and any attempt to think of things as having a persistent essence—even if we understand it as an observed pattern rather than an ideal form—is fundamentally unsupported by our assumption of the primacy of intentionality.

That being said, we can still talk about things as things because our naive idea of thingness is not useless, only fuzzy. So long as we remain clear that our demarcation of things is an ephemeral choice, albeit an ephemeral choice based on our observations of reality, we may talk about things and their existence over time. Thus, to return to our reduction of self experience, we can say that the same I is both the subject and object of the phenomenon of self experience so far as we understand the I to be the same thing before and after each particular experience. That the I may change so much over enough such experiences that we would no longer recognize the I as the same I given a long enough period of time is a matter of our choices about when a thing stops being itself and starts being something new—the same choices we must always make in building ontology.

Feedback

We now have only one question remaining to address with our reduction: is the I experiencing itself as itself or is it only experiencing itself? That is, when we say that {I, experience, I} do we mean that I experiences itself as an ontological thing or do we mean only that the I experiences itself as an ontic thing without subjective identification of the object as itself? Our experience of consciousness suggests that we mean the former, but understanding clearly its relationship to the latter will, I think, help pull together the concepts we've just discussed and prepare us for a deeper discussion of consciousness, qualia, and noematology.



To avoid confusion with consciousness, let's begin by considering the self experience of something generally not considered conscious—a steam engine with a governor. If you're not familiar with steam engines, the operation of one with a governor is straightforward: the boiler heats water to produce steam, the steam applies pressure to rotate the turbine, the turbine turns the drive shaft, part of the drive shaft's mechanical energy is used to spin the governor, and the governor is connected to a throttle valve controlling the amount of steam that reaches the turbine. When the engine starts the throttle is open, so steam pressure increases, turning the turbine faster, spinning the governor faster, closing the throttle a little. This leads to decreasing steam pressure, turning the turbine slower, spinning the governor

slower, opening the throttle a little, which causes increasing steam pressure, turning the turbine faster, spinning the governor faster, closing the throttle a little, and on and on until the engine runs out of heat or water, thus making sure the steam engine's drive shaft always turns at a nearly constant speed. We call this a homeostatic or negative feedback process.

What's phenomenologically interesting about this steam engine is that it has no idea what it's doing. As far as we can tell it doesn't know what it is or how it works, i.e. it has no ontological self awareness, nor does it encode the idea that the drive shaft should turn at a particular speed, yet because the steam engine exists as an ontic thing we can say that the steam engine experiences itself because the governor gives the engine a way to experience past experiences of its own ontic existence giving the steam engine the appearance of telos. The phenomena {slower steam engine, if too fast, steam engine} and {faster steam engine, if too slow, steam engine} arise because the governor connects the output of the steam engine to its input, and in doing so creates an ontic thing that, through our ontological lens, appears to persist, have self experience, and regulate its own action over time. This makes the governor less than conscious but more than inert: it makes it cybernetic.

The term "cybernetic" has been abused a lot, so to be clear when I say the governor is cybernetic I precisely mean that it's an (ontic) thing that experiences its (ontic) self. Unfortunately I have to heap my own abuse onto "cybernetic" because, as we've defined it, everything worthy of thingness must be cybernetic. Consider a seemingly inert, non-cybernetic thing, like a rock. Rocks appear entirely inactive, yet rocks manage to stay together and be things despite the universal tendency towards entropy. The stuff of a rock is constantly interacting to maintain itself rather than merge into a soup of its surroundings, and so it is that rocks, after our understanding, must be experiencing themselves, and so must be cybernetic, and similarly all things, to the extent that they are things and differentiate themselves within the world, are cybernetic.

Cybernetics is still useful as a concept though because there is a sense in which a steam engine is more cybernetic than a rock. The self experience of the steam engine generates more thermodynamic entropy (in the form of exhaust heat) than the self experience of the rock (in the form of diffusion), and so the steam engine's self experience can, though doesn't necessarily have to, contain more information than the self experience of the rock. Indeed, the rock is perhaps a more efficient self experiencer than the steam engine since the rock's self experience makes almost no information by also generating almost no entropy, but we are more interested in the steam engine because it produces a whole bit of information—whether to open or close the throttle—by creating lots of heat and mechanical energy. And cybernetic things only get more interesting as they produce more information.

Computers, for example, produce so much information that they create things out of information. I'm typing this on a computer, and as a result the computer is creating a thing we call a document that consists purely of information within the computer system. Indeed, this document has no physical representation of its own: it exists purely within the self experience of the computer due to the complex interactions of its many parts, and there's no way to separate the document from the computer other than to produce another document in another computer or to print a representation of it on paper. And if cybernetic things can contain things within the information of their self experience, what happens if these information things start to experience themselves?

Well, for one it means that information things, like the document, have ontological existences separate from their ontic existences—you know, the map is not the territory—because we can create many instances of the "same" document yet they will all be made up of different information being produced by different computers. It also means that ontological things are manifested as ontic things, so ontological things can also experience themselves via the ontic, which is to say they are embodied, and when ontological things experience themselves we say that the ontic things they exist in are phenomenally conscious rather than merely cybernetic because they both experience themselves and

experience themselves as themselves, as ontological things. Thus the I of self experience both experiences itself and experiences itself as itself, answering our final question.

At last our reduction is, for now, complete. Through it we've seen that:

- Things exist ontologically as patterns within our experience of them.
- Things exist ontically as clusters of stuff within the world.
- Things exist ephemerally though chains of experiences creating the perception of time.
- Through ephemeral existence over time, things can feed back experiences of themselves to themselves, making them cybernetic, and in so doing create information.
- Things can exist within information, those things can experience themselves, and it's from those information things that ontology, and thus consciousness, arises.

Next time, in our final installment before getting back to AI alignment, we'll tackle questions about consciousness, qualia, and how the contents of qualia combine. You can read it here.

Introduction to Noematology

This is a linkpost for https://mapandterritory.org/introduction-to-noematology-fac7ae7d805d

NB: Originally posted on Map and Territory on Medium, so some of the internal series links go there.

Last time we performed a reduction of the phenomenon of conscious self experience and through it discovered several key ideas. To refresh ourselves on them:

- Things exist ontologically as patterns within our experience of them.
- Things exist ontically as clusters of stuff within the world.
- Things exist ephemerally though chains of experiences creating the perception of time.
- Through ephemeral existence over time, things can feed back experiences of themselves to themselves, making them cybernetic, and in so doing create information.
- Things can exist within information, those things can experience themselves, and it's from those information things that ontology, and thus consciousness, arises.

We covered all of these in detail except the last one. We established that the feedback of things created from the information of feedback gives rise to ontology by noting that information things have ontological existences that transcend their ontic existences even as they are necessarily manifested ontically. From there I claimed that, since people report feeling as if they experience themselves as themselves, consciousness depends on and is thus necessarily created by the ontological experience of the self. Unfortunately this assumes that our naive sense of self could not appear any other way, and in the interest of skepticism, we must ask, can we be sure there is not some more parsimonious way to explain consciousness that does not depend on ontological self experience?

I think not, but some philosophers disagree. Consider the idea of p-zombies: philosophical "zombies" that are exactly like "real" people in all ways except that they are not really conscious. Or consider John Searle's Chinese room, where a person who cannot understand Chinese nevertheless is able to mimmic a native Chinese speaker via mechanistic means. In each of these cases we are presented with a world that looks like our own except that consciousness is not necessary to explain the evidence of consciousness.

Several responses are possible, but the one I find most appealing is the response from computational complexity. In short, it says that p-zombies and Chinese rooms are possible by unrolling the feedback loops of ontological self experience, but this requires things that we think are conscious, like people, to produce exponentially more entropy, be exponentially larger, or run exponentially slower than what we observe. Given that people are not exponentially hotter, larger, or slower than they are, it must be that they are actually conscious. Other arguments similarly find that things that theoretically look like conscious entities while not being conscious are not possible without generating observable side-effects.

So if it is the case that reports of feeling as though consciousness includes experiencing the self as the self describe a necessary condition of consciousness, then

ontological self experience must be necessary to consciousness. This is not to say it is a sufficient condition to explain all of consciousness, though, since that would require explaining many details specific to the way consciousness is embodied, so we properly say that ontological self experience explains *phenomenal* consciousness rather than the phenomenon of consciousness in general. Nevertheless there is much we can do with our concept of phenomenal consciousness that will take us in the direction of addressing AI alignment.

Qualia and Noemata

To begin, let's return to our reduction of {I, experience, I} in light of our additional understanding. We now know that when we say "I experience myself" we really mean "I experience myself as myself", so it seems our normalized phenomenon should be {I, experience, I as I}. We could have instead written this as {I, experience as I, I} since it is through self experience that the I sees the ontic self as an ontological thing, but the former notation is useful because it exposes something interesting we've been assuming but not yet explored: that the subject of a phenomenon can experience an ontological thing as object. Yet how can it be that a thing that exists only within experience can become the object of experience when experience happens between two ontic things?

The first part of the answer you already know: the ontological existence of a thing necessitates ontic manifestation. Like with the computer document, a thing might have an ontological existence apart from its ontic existence, but ontological existence implies ontic existence since otherwise there is no stuff to be the object of any experience, and for it to be otherwise would be to suppose direct knowledge of ontology, which we already ruled out by choosing empiricism without idealism. Thus an ontological thing is also an ontic thing, the ontic thing can be the object of experience, and so the ontological thing can be the object of experience. But only understanding that ontological things can be the object of experience in this way fails to appreciate how deeply ontology is connected to intentionality.

Notice that in order to talk about a phenomenon as an intentional relation we must identify the subject, experience, and object. That is, we, ourselves phenomenological subjects, see a thing that we call subject, see a thing that we call object, and see them interacting in some way that we can reify as a thing that we call an experience, i.e. we see the members of the intentional relation ontologically. If we don't do this we fail to observe the phenomenon as an intentional relation and thus as a phenomenon, because if we fail to see the phenomenon as ontological thing we have no knowledge of it as a thing and can only be affected by it via direct experience of the ontic in the same way rocks and trees are affected by phenomena without knowing they exist or are being affected by experiences. This means that the object of experience, insofar as the subject can consider it the object of experience, has ontological existence by virtue of being the object of experience, even if that ontological existence is not or cannot be seen by the subject of the experience. Thus of course "I as I" can be the object of I's experience because we have already proved it so by considering the possibility that it is.

So if "I as I" can be the object of the I's experience of itself, why bother to think of the phenomenon this way rather than as {I, experience as I, I}? As way of response, consider how the I comes to have ontological existence: the I experiences the ontic I, this creates a feedback loop of experience over time that allows the creation of information, and then an ontic thing that the I can experience emerges from that

information. That information-based, ontic thing carries with it the influence of the ontic I—just as the bit expressing the state of the throttle in the steam engine is created via the governor's feedback loop and carries with it the influence of the reality of the steam engine's configuration—so it is causally linked to the I's ontic existence. We then call this "I as I" because it is the thing through which the I experiences itself as a thing by the I making the phenomenon of experience of the self as ontic thing the object of experience. Thus by thinking of the phenomenon of self experience as {I, experience, I as I} we see it has another form, namely {I, experience, {I, experience, I}}.

This highlights the structural difference between consciousness and cyberneticness. A cybernetic thing, as far as it is cybernetic, only experiences its ontic self directly via its feedback loops over itself. A conscious thing, though, can also experience its ontic self indirectly through feedback loops over the things created from the information in its cybernetic feedback loops. It's by nesting feedback loops that the seed of phenomenal consciousness is created, and we give the things created by these nested feedback loops and the phenomena that contain them special names: noemata and qualia, respectively.

"Noema" is Greek for both thought and the object of thought, and the project of phenomenology was started when Husserl saw what we have now seen by building on Brentano's realization that mental phenomena, also known as qualia, are differentiated from other physical phenomena by having noemata as their objects. That noemata are phenomena themselves, and specifically the phenomena of cybernetic things, was to my knowledge first well understood by Hofstadter, although Dretske seems to have been the first to take a stance substantially similar to mine, because it required the insights of control theory and the other fields that make up cybernetics to understand that mental phenomena are not something special but a natural result of nested feedback. This is also why early phenomenologists, like Husserl, tended towards idealism, while later ones opposed it: without an understanding of cybernetics it was unclear how to ground phenomenology in physical reality.

To reach the fairly unorthodox position I've presented here, it took an even deeper knowledge of physics to trust that we could make intentionality as central to epistemology as we have and maintain an existentialist stance. As such you may be left with the feeling that, while none of what I have presented so far is truly novel, I have left unaddressed many questions about this worldview. Alas my goal is not to provide a complete philosophical system but address a problem using this philosophical framework, so I have explained only what I believe is necessary to that end. We move on now to less well trod territory.

Noematology

Having identified noemata as the source of consciousness, our view of consciousness is necessarily noematological, i.e. it is based on an account of noemata. This invites us to coin "noematology" as a term to describe our study of the phenomena of consciousness through the understanding of noemata we have just developed. It also conveniently seems little used and so affords us a semantic greenfield for our technical jargon that avoids some of the associations people may have with related terms like "qualia", so we take it up in the spirit of clarity and precision.

Noematology, despite being newly minted, already contains several results. The first of these follows immediately from the way noemata arise. Noemata, being simply the result of nested feedback, appear everywhere. That is to say, noemata are so pervasive that our theory of phenomenal consciousness is technically panpsychic. Specifically, since all things are cybernetic, all things must also contain in their self experiences information out of which things emerge, and those information things must themselves be cybernetic insofar as they are things, thus they are noemata, hence all things must be phenomenally conscious. Of course not all things are equally phenomenally conscious just as not all things are equally cybernetic: some things produce more and more used noemata than others, just as some things produce more and more used information than others. Ideas like integrated information theory act on this observation to offer a measure of consciousness that let's us say, for example, that mammals are more conscious than trees and that rocks have a consciousness measure near zero. Integrated information theory also shows how the panpsychism of phenomenal consciousness is vacuous: it's true, but only because it pushes most of the things one might like to claim via panpsychism out of the realm of philosophy and into the realm of science and engineering. Put another way, consciousness may be everywhere in everything, but it's still hard to be conscious enough for it to make much of a difference.

That the manifestation of consciousness, especially the consciousness of things like humans, is complicated gives purpose to noematology because it helps us see insights that are normally occluded by implementation details. For example, that noemata are created by the nesting of feedback loops within feedback loops immediately implies the existence of meta-noemata created by the nesting of feedback loops within noemata. And if this nesting can be performed once, it can be performed many times until there is not enough negentropy left to produce even one bit of information from an additional nesting. These multiple orders of noemata can then be used to explain the qualitative differences observed during human psychological development, and that higher-order noemata, which we might also call the expression of higher-order consciousness, are necessary to create qualia like tranquility and cognitive empathy, but these topics are beside our current one. For now we turn our attention to the relationship between noemata, axias, and ethics because it will ground our discussion of AI alignment.

Axiology, Ethics, and Alignment

Philosophy is composed of the study of several topics. Naturally, there is some disagreement on what those topics are, what to call them, and how they relate, but I tend to think of things in terms of epistemology, ontology, and axiology—the study of how we know, the study of what we know, and the study of why we care. All three are tightly intertwined, but if I had to give them an ordering, it would be that epistemology precedes ontology precedes axiology. That is, our epistemological choices largely determine our ontological choices and those in turn decide our axiological choices. Thus it should come as no surprise that I had to address epistemology and ontology before I could talk about axiology.

Of course the irony is that we actually investigate philosophy the other way around because first we ask "why?" by wanting to know, then we ask "what?" by knowing, and only finally can we ask "how?" by considering the way we came to know. The map, if you will, is drawn inverted relative to the orientation of the territory. So in some ways we have been studying axiology all along because axiology subsumes our founding question—why?—but in other ways we had to hold off talking about it until

we had a clear understanding of how and what it means to ask “why?”. With that context, let’s now turn to axiology proper.

Axiology is formally the study of axias or values just as ontology is the study of ontos or being and epistemology is the study of episteme or knowledge. An axia then is something of value that we care about, or put another way, since it’s the object of a phenomenally conscious experience of caring, it’s a noema to which we ascribe telos or purpose, so we might think of axiology as teleological noematology, but to bother to think of something is to give it sufficient telos that it was thought of rather than not, so in fact all noemata we encounter are axias by virtue of being thought of. Non-teleological noemata still exist in this view, but only so long as they remain unconsidered, thus for most purposes noematology and axiology concern the same thing, and the choice of which term to use is mostly a matter of whether we wish to emphasize traditional axiological reasoning or not.

To make this concrete, consider the seemingly non-teleological, value-free thought “this is a pancake”. Prior to supposing the existence of the pancake there could have been a thought about the pancake which was valueless because it existed but was not the object of any experience, but as soon as it was made object it took on purpose by being given the role of object in an intentional relation by the subject experiencing it. From there the subject may or may not assign additional purpose to the thought through its experience of it, but it at least carries with it the implicit purpose of being the object of experience. As with thoughts of pancakes, so too with all thoughts, thus all thoughts we encounter are also values.

Within this world of teleological noemata we now consider the traditional questions of axiology. To these I have nothing special to add other than to say that, when we take noemata to be axias, most existing discussions of preferences, aesthetics, and ethics are unaffected. Yet I am motivated to emphasize that noemata are axias because it encourages a view of axiology that is less concerned with developing consistent systems of values and more concerned with accounts that can incorporate all noemata/axias. This is important because the work of AI alignment is best served by being maximally conservative in our assumptions about the sort of conscious thing we need to align.

For example, when working within AI alignment, in my view it’s best to take a position of moral nihilism—the position that no moral facts exist—because then even if it turns out moral facts do exist we will have built a solution to alignment which is robust to not only uncertainty about moral facts but to the undesirability of moral facts. That is, it will be an alignment solution which will work even if it turns out that what is morally true is contrary to human values and thus not what we want an aligned AI to do. Further, if we assume to the contrary that moral facts do exist, we may fail to develop a sufficiently complete alignment solution because it may depend on the existence of

moral facts and if we turn out to be mistaken about this such a solution may fail catastrophically.

Additionally, we may fail to be sufficiently conservative if we assume that AI will be rational or bounded-rational agents. Under MIRI's influence the assumption that any AI capable of posing existential risk will be rational has become widespread within AI safety research via the argument that any sufficiently powerful AI would instrumentally converge to rationality so that it does not get Dutch booked or otherwise give up gains, but if AGI were to be developed first using machine learning or brain emulation then we may find ourselves in a world where AI is strong enough to be dangerous but not strong enough to be even approximately rational. In such a case MIRI's agent foundations research program might not be of direct use because it makes too strong of assumptions about how AI will reason, though it would likely offer useful inspiration about how to align agents in general. In the event that we need to align non-rational AI, addressing the problem from axiology and noematology may prove fruitful since it makes fewer assumptions than decision theory for rational agents.

Even if we allow that an AI capable of posing an existential threat would be rational, there is still the axiological question of how to combine the values of humans to determine what it would mean for an AI to be aligned with our specific values. To date there have been some proposals, and it may be this problem can be offloaded to AI, but even if we can ask AI to provide a specific answer we still face the metaethical questions of how to verify if the answer the AI finds is well formed and how to ensure the AI will find a well formed answer. In view of this we might say that alignment asks one of the questions at the heart of metaethics—how do we construct an ethical agent?—and solving AI alignment will necessarily require identify a "correct" metaethics. In this case the study of AI alignment is inseparable from axiology and, in my view, noematology, so these are important lenses through which to consider AI alignment problems in addition to decision theory and machine learning.

These are just among some of the topics for which we wish to address with our noematological perspective. And there are of course many topics outside AI alignment on which it touches, some of which I have already explored on this blog and others which I have considered only in personal conversations or during meditation. I will have more to say on these topics in the future, especially as they relate to AI alignment, but for now this completes our introduction to existential phenomenology and noematology. On to the real work!

AI Alignment and Phenomenal Consciousness

This is a linkpost for https://mapandterritory.org/ai-alignment-and-phenomenal-consciousness-2ca23de6aebd

NB: Originally posted on Map and Territory on Medium, so some of the internal series links go there.

In the initial feedback I've received on my attempt to formally state the AI alignment problem, the primary objection I've heard is that I assume any AI worth aligning will experience qualia. In particular, some think we have to worry about aligning AI that is cybernetic but not phenomenally conscious. I think they are mistaken in thinking cybernetic-only AGI is possible, but in all fairness this is probably due to a failure on my part to adequately explain phenomenal consciousness, so I'll try here to make clearer the case for phenomenally conscious AGI and how this relates to AI alignment.

To begin, though, let's step back and consider what we mean by "artificial intelligence"? The artificial part is easy: we mean something constructed through a deliberate design effort rather than arising naturally, i.e. it is the handiwork of conscious things rather than the outcome of an unconscious process. Thus artificial intelligence stands in contrast to natural intelligence, like the intelligence of animals that arose through evolution. The intelligence part is harder because by intelligence we mean multiple things: possession of goal directed behavior (telos), ability to model the world (ontology), and an ability to combine telos and ontology to come up with new solutions to problems (find undiscovered algorithms). Generally we might say intelligence is something like an ability to systematically increase local informational complexity—optimize the nearby parts of the world—by increasing global entropy. Taken together then, artificial intelligence and artificially intelligent agents can be said to be designed optimization processes.

This means, of course, that things as simple as steam engines are a kind of artificial intelligence even if they aren't especially intelligent since a steam engine is increasing global entropy in the form of waste heat in order to produce mechanical power. And even if we put a governor on our steam engine it is still only cybernetic—a thing that experiences itself—and not phenomenally conscious—a thing that experiences itself experiencing itself—but, as we'll see, it doesn't take a lot to make our steam engine jump into the realm of phenomenal consciousness.

The steam engine with a governor, as I've previously explained, is a cybernetic thing that produces at least one bit of information about whether the throttle is open or closed. Although it would be needlessly complex, suppose we added a governor onto the governor to regulate how quickly the governor adjusts the throttle. This governor governor has a simple operation: if the throttle was open within the last second, it doesn't allow the throttle to close and vice versa. In doing so it creates a kind of memory for the steam engine about the state of the throttle, generates ontology by interpreting and representing the state of the throttle within the governor governor, and experiences itself experiencing itself through the governor governor experiencing the governor experiencing the steam engine. These features all imply this modified steam engine is phenomenally conscious.

To consider an example closer to the edge of our capabilities in artificial intelligence, creating phenomenally consciousness with machine learning is also trivial. A simple machine learning algorithm is a cybernetic process that iterates over data to produce a non-cybernetic model (the model is not cybernetic because it is essentially a lookup table or function that does not experience itself). More complex machine learning algorithms can produce cybernetic models with memory and, depending on the implementation of the algorithm, this can make the algorithm phenomenally conscious. If a machine learning algorithm generates cybernetic models that generate cybernetic models or self improves, then it's solidly in the realm of phenomenal consciousness. After all, the minimum requirement for phenomenal consciousness is little more than a loop nested inside another loop!

But just because phenomenal consciousness is easy to place in a system and is in use in today's leading edge of AI development, we don't necessarily want to create phenomenally conscious AI. Eliezer Yudkowsky, for example, has argued strongly against creating conscious AI for ethical and technical reasons, and although he was referring to a naive sort of consciousness rather than phenomenal consciousness, it would be safer to make AI that are less capable rather than more, so we only want to create phenomenally conscious AI if it's necessary to our ends. Unfortunately, I think if we want to create AGI—artificial general intelligence—it will necessarily be phenomenally conscious.

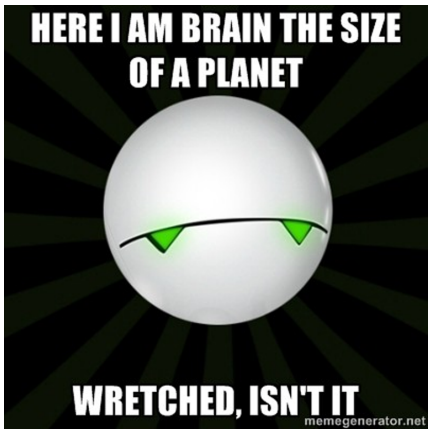
Let's step back again and consider what we mean by the "general" in AGI. "General" stands in opposition to "narrow" in AI where narrow AI are designed optimization processes that work only in one or a few domains, like chess or language translation. Artificial general intelligence, on the other had, is expected to work in arbitrarily many domains, including domains the AI has not been trained on, because it could presumably train itself the way humans can when it encounters novel scenarios or otherwise adapt to new situations. It might fail at first, but it will learn and grow in capability as it addresses a broader space of experiences. It is only such general AI that I posit must be phenomenally conscious, especially since we know by their existence that cybernetic-only narrow AI is possible.

Suppose we could create a cybernetic-only AGI. Such a thing, being not phenomenally conscious, would necessarily have no ontology or ability to model the world, so it would be a kind of philosophical zombie that behaves like a phenomenally conscious thing but is not. P-zombies are possible, but they are not cheap, with a p-zombie requiring exponentially more computational resources than a behaviorally equivalent phenomenally conscious thing in terms of the number of cases it would be expected to handle (Update 2018–03–20: I try to prove this formally). That is, a behaviorally equivalent p-zombie needs a separate cybernetic system to handle every situation it might find itself in because it can't model the world and must have its "ontology" hard coded. So, if we could create a cybernetic-only AGI, how big would it be, both in terms of cybernetic subsystems and volume? My Fermi estimate:

- There are on the order of 10,000 unique words needed to fully express ideas in any given human language.
- A sentence in a human language has on the order of 10 words.
- Assuming every sentence of 10 words in any language describes a unique scenario, a human-level AGI must handle at least $10{,}000^{10}=10^{40}$ scenarios.
- Since 3 levels of recursion are enough for anyone, let's conservatively suppose this means scenarios interact to create at least $(10^{40})^3=10^{120}$ situations an AGI must deal with, requiring 10^{120} cybernetic subsystems.

- AlphaZero can train to handle a new situation, like a new game, in the order of 1 hour, but to be conservative and assume future improvements let's assume our AGI can train on a scenario in 1 minute.
- That means our AGI needs to perform 10^{120} minutes, or $\sim 1.9 \times 10^{114}$ years, worth of training to build models to handle all the scenarios it needs to be general.
- Supposing we are willing to take on the order of 10 years to build our AGI and we can parallelize the training over that period, building an AGI would require $1.9 \times 10^{114}/10 = 1.9 \times 10^{113}$ computers each on the scale of AlphaZero.
- It's unclear what AlphaZero's computational needs are, but AlphaGo Zero apparently runs on only a single server with 4 TPUs. Let's conservatively assume this means we need 1U of rack space or $\sim 15{,}000 cm^3$ or $\sim 0.015 m^3$ to train a model to handle each scenario.
- So to get an AGI you need $1.9 \times 10^{113} \times 0.015 m^3 = 1.85 \times 10^{111} m^3$ of compute, not leaving space for cooling, power, etc.
- The Earth has a volume of $\sim 1.1 \times 10^{21} m^3$, so our AGI would require 1.68×10^{90} Earths worth of computers.

I'm sure we could make these calculations more accurate, but that's not the point of a Fermi estimate; the point is to show the scale of building an AGI as capable of a human that is also a p-zombie, even if my calculations are wrong by several orders of magnitude. If we tweaked the numbers to be maximally favorable to building p-zombies, taking into account improvements in technology and the problem being easier than I think it is, we would still end up with needing more than an entire Earth's worth of computers to do it. Building a human-level p-zombie AGI would be asking for a planet-sized brain, and we know how that would turn out.



Jokes aside, I don't go through this exercise to poke fun at the idea we could build an AGI that is not phenomenally conscious. My objective is to stress that any practical AGI project will necessarily be looking at building something phenomenally conscious because it's the only way to fit the amount of complexity needed into a reasonable amount of resources. I don't think people working on AI capability are confused about this: they know that giving systems what I call phenomenal consciousness allows them to do more work with less resources, and doing this seems to be the direction they will naturally go, even with narrow AI, as cybernetic-only solutions become prohibitively expensive to improve. But if AI safety researchers were hoping for cybernetic-only AGI, it alas seems it will definitely be phenomenally conscious.

That said, seed AI might allow us to avoid phenomenally conscious AGI for a while. The idea of seed AI is to first create a simple AI system that will bootstrap itself into a more powerful one by improving itself or designing its successor. Maybe we could

design seed AI that is cybernetic-only and let the seed AI take over the responsibility of designing phenomenally conscious AGI. In such a scenario, might we need to consider the alignment of cybernetic-only AI?

In short, no. We would be interested in designing a seed AI such that it used the results of alignment research so that any more powerful, phenomenally conscious AGI it created would be aligned, but the seed AI itself would not need alignment because, to make a broader point, there is a sense in which something that is not phenomenally conscious cannot be aligned because it does not value anything because it doesn't know what anything is. To put it another way, we can't build tools—things we use for a particular purpose that are phenomenally unconscious—that cannot be misused because they lack the complexity necessary to even notice they are being misused, let alone do something to avoid that.

To give an example, a crowbar lacks the complexity to know if it's being used to open a crate or break into a building, much less know if opening the crate or breaking into the building is "good" or "bad". Yes, we could make fancy, cybernetic crowbars with tiny computers that noticed what they were being used for and would stop working when they detected a "bad" scenario, but being cybernetic-only it would not be general and only able to handle situations it was trained to recognize. Use the crowbar in a novel situation and it may promptly become "unaligned" because it was never "aligned" in the first place: it just did what it was designed to do, even if that design included complex, narrow AI that worked in a lot of cases. If you want to build aligned things, AGI or otherwise, you have to make them phenomenally conscious because that's the only way the thing can possibly share the operator's values in general.

I hope this makes clear why I think AGI will be phenomenally conscious and why AI alignment is a problem about phenomenally conscious agents. I invite further feedback on developing these ideas, so please comment or reach out with your thoughts, especially if you disagree.

Computational Complexity of P-Zombies

This is a linkpost for https://mapandterritory.org/computational-complexity-of-p-zombies-fc56909af96f

I've claimed that p-zombies require exponentially more resources than phenomenally conscious agents and used this to justify that AGI will necessarily be phenomenally conscious. I've previously supported this claim by pointing to a related result in integrated information theory showing that p-zombies constructed as feed-forward networks are exponentially larger than behaviorally equivalent conscious networks, but now I'd like to see if I can get a similar result in noematology.

NB: This account is provisional and not sufficiently well-founded to constitute a proof given my assumptions. That said, it should be precise enough to be falsifiable if wrong.

First, some notation. Given an agent A, let a **behavior** be a pair {x, A(x)} where x is the world state before the agent acts and A(x) is the world state after. For example, if when holding a teacup A takes a sip of tea, then x is "holding a cup of tea" and A(x) is "taking a sip of tea". If we have a second agent, Z, we may wish to ask if, given the same initial world state x, Z(x)=A(x), viz. Z and A exhibit the same behavior when presented with x. In an obvious sense this question is meaningless because A and Z cannot be said to observe the exact same reality nor is there to any epistemically valid method within phenomenalism by which we may directly compare two world states, but given an observer agent O, we can consider if A(x)=Z(x) with respect to O's experience of A(x) and Z(x) and this is what we will mean when we ask if A exhibits the same behavior as Z.

Now let A be a phenomenally conscious agent and Z a p-zombie of A, meaning that Z and A cannot be distinguished based on observed behavior and Z is not phenomenally conscious (although it is, of course, cybernetic). More formally, we are supposing that A and Z agree on the set of behaviors X such that for all {x, y} in X, y=A(x)=Z(x). I then claim that Z is exponentially larger than A with the cardinality of X, where by "larger" I mean physically made of more stuff.

The short explanation of my reasoning is that Z must do with only the ontic what A does with ontology. Unfortunately this hides a lot of my understanding about what phenomenal consciousness does and how it works, so instead I'll explain by relating A and Z to computational models that can describe them and use those to show that Z must be exponentially larger than A.

Since Z is not phenomenally conscious it cannot create information within itself because it has no feedback process with which information could be created (if it did it would then be phenomenally conscious and no longer a p-zombie). Computationally this describes the class of computers we call deterministic finite automata or DFAs. A, on the other hand, can create information via feedback; this gives it "memory" which we can model as a Turing machine with a finite tape (the tape must be finite since it is embodied in the real world rather than a theoretical construct). Other models are possible, but these should suffice for our purposes.

Since Z is a DFA, it must have at least one state for each world state x it encounters to ensure the correct mapping from x to y for each {x, y} in X, thus Z has at least O(|X|) states. This puts a lower bound on the size of Z, so to show that Z is exponentially larger than A in terms of |X| we need to show that A has an upper bound on its size of O(lg|X|) states and tape length. We will do this by constructing a phenomenally conscious version of Z by converting it from a DFA to a Turing machine. It's easier to understand how to convert a finite Turing machine into a DFA, though, so we'll do that first and then reverse the process to show that A is no larger than O(lg|X|).

Since A is a finite Turing machine it can be converted into a DFA by an algorithm analogous to loop unrolling that requires creating at least one state in the DFA for each possible configuration of the tape. Let us denote this DFA-ized version of A as A'. This means that, given that A has a tape of length n, we need A' to contain at least 2^n states. Supposing A consists of a state machine of fixed sized independent of |X| and always encodes information about each world state x on the tape, then A will need a tape of at least length lg|X| to encode every x, and A' will need to contain at least 2^(lg|X|)=|X| additional states to replace the tape. This gives us a lower bound on the size of Z since A' has O(|X|) states, so Z has O(|X|) states.

Reversing this algorithm we can construct a Turing machine Z' from Z by using the tape to encode the O(|X|) states of Z on at least O(lg|X|) of tape where Z' otherwise consists of a constant-sized finite state machine that emulates Z. Since Z' is a Turing machine it meets our requirements for phenomenal consciousness and so gives a lower bound on the size of A. A has a natural upper bound of O(|X|) since A can be simply constructed as Z with a tape, so A has between O(lg|X|) and O(|X|) states and tape length.

This unfortunately does not give us the desired result that A has an upper size bound of O(lg|X|) and only says that A is no larger than Z and Z is at most exponentially larger than A in terms of |X|. Nonetheless I believe it to be true that A also has at most O(lg|X|) states and tape length because this is what we empirically see when constructing, for example, less phenomenally conscious state machine solutions to programming problems rather than more phenomenally conscious versions using nested loops or recursion. I also suspect it's possible to prove an upper bound lower than O(|X|) closer to O(lg|X|) for A but have not found such a proof in the literature nor proved this myself. There may alternatively be a weaker relationship between the sizes of A and Z that has the same consequence that Z must be much larger than A but is quantitatively less than exponential.

(If you're interested in collaborating with me on this please reach out since this would be a nice result to have locked down for when I publish my work on formally stating the AI alignment problem! I believe it's possible, but I'm rusty enough on automata theory that the effort I would have to put in the prove this is high. If you already have automata theory fresh in your mind then this is likely more straightforward.)

Finally we need to connect this notion of number of states and tape length to physical size. This is straight forward, but to be explicit about it each state or tape position needs to be embodied in physical stuff. Even assuming a single transistor or a single bit of storage is all that is needed for each state or tape position, this means we can use our size calculations to estimate and compare the relative physical sizes of A and Z by setting up a correspondence between theoretical state and tape constructs and physical computer implementations. This let's us conclude at least a weak version of what we set out to prove: that a p-zombie Z of a phenomenally conscious agent A is no smaller than A in terms of |X| as measured in physical stuff, is probably much

larger than A, and is exponentially larger than A if we can prove that A has at most O(lg|X|) states and tape length.