Consciousness Actually Explained: EC Theory

CASEY
JAN 08, 2023

(Preface/note on usage: Even if you feel well acquainted with this topic, I would still urge you to read through the introductory sections, and to steel yourself for framings/phrasings that normally would trigger you to stop reading. This problem has remained unsolved for so long in part due to improper framings of the problem itself. Don't interpret any lack of reference to traditional theories as ignorance; I address them near the end.)

*(**Preface for LessWrong readers specifically**: This is a complete and reductive solution to a conceptual problem. The moment you feel like I've taken a left-turn into "woo", consider that I've considered this very possibility and forewarned you against concluding I will resort to magic, or something like time-cube-esque theorizing. If anything this is exactly the kind of document that neuroscience and AI folks would want; to get the philosophers to stop pestering them about the "hard problem" so they can continue their practical research as usual.*

> [Addressing Traditional Theories - Throwing Down the Gauntlet]

Defining the Problem and Premises

"Consciousness" as a word can mean different things, such as being merely awake as opposed to "unconscious" or asleep.

The meaning of consciousness discussed here is **subjective internal experience**, or the quality of there being "something **that it is like** to be something". To get a grip on this concept, we distinguish consciousness from the contents of consciousness. This content could be anything that fills our experience or awareness. For example we can be aware of the traditional 5 senses and things we detect within them (ex: sight of a red apple), emotions, indescribable thoughts and other mental events, memories, hallucinations/dreams, etc. These are all possible contents of consciousness, but consciousness itself is just the fact that there is an experience at all of any particular content; that there is subjective experience at all for any particular thing that could be experienced.

The classic example is to imagine the consciousness - the internal subjective experience - of a bat. Bats share most aspects of the 5 traditional senses that humans have, but they also have the additional sense of sonar. What is it like to be a bat? We don't know, as we aren't a bat and don't have a complete understanding of consciousness. But we can reasonably guess there is *something* that it is like to be a bat, however alien that experience would be, given their additional sense of sonar. In the case of humans, we don't have to guess. There is clearly **something that it is like** to be a human; we have internal experience; there is an "inner-reality"; "the lights are on" inside.

The example of the hypothetical consciousness of a bat is useful in at least one way: it highlights that while the contents of consciousness could be anything at all - even something truly alien to humans like sonar - the concept of consciousness or experience itself is the same, whether applied to humans, bats, or anything else we are guessing may have consciousness.

We must also not confuse our mission by introducing extraneous and unnecessary concepts which often pop-up when discussing philosophy of mind. Unless these later

emerge as relevant to our very specific mission of explaining just consciousness - we are **not** talking about, concerned with, or going to try to explain:

- Free will or intentionality, or a lack of either
- Emotions or dispositional states
- Selfhood or personhood
- Having a "locus" of consciousness; how our experiences seem somewhat unified
- The nature of pleasure, wellbeing, pain, suffering, good, evil, etc.
- How we can speak, think, hypothesize, move, be intelligent, etc.
- How we can have any particular content of consciousness, even including how we can have "internal-access" or "self-reflection": the ability to know and internally reflect upon the activity and state of our body or mind. All this is just more content. (example: Cows are likely conscious, but probably don't have the ability to self-reflect on their emotions or mental state. And while they are aware of their body, we are concerned with awareness itself, not that they are aware of any particular content, such as the senses that indicate to them that they have a body.)

Many of those bullets can also be reframed into what is traditionally known as the distinction between the "hard" problem vs "easy" problems of consciousness. Here "easy" problems are imagined to be those that can be eventually explained merely by specifying a mechanism that could accomplish that action. An example would be the capability of sight, which we could in principle eventually explain by describing the physical events of photons entering a retina/camera, sending a signal to the brain/cpu, and being transformed into some medium of representation that a program/brain could integrate into a model of the outside world. We are already building such robots, and while we don't know the precise mechanism or procedure by which the brain does this, there is no deep mystery here. It is an ongoing problem of engineering and neuroscience.

The "hard" problem is the problem of how anything at all - any process, program, complex object, algorithm, etc. - could be conscious. Beyond mere function - even complex functions that could explain the ability to intelligently act/speak/move in the world - how and why is there an "inner-reality" that accompanies this functioning? Why

is there an experience of anything at all? Surprisingly, not everyone shares this intuition that there is a distinction between the hard vs easy problems, or that there is a hard problem to solve at all. For those still unconvinced, I would direct them to these 2 very brief articles by Sam Harris:

https://samharris.org/the-mystery-of-consciousness/

https://samharris.org/the-mystery-of-consciousness-ii/

If you remain unconvinced that there is a hard problem (which isn't to believe that this problem is unsolvable!), or are unable to at least conceptualize what the supposed difficulty or distinction being discussed is, there is still value in you continuing to read this explanation. Let's establish those who think there is a hard problem/question as "hard-questioners", and those who don't as "non-hard-questioners". It's important to note that you can remain a hard-questioner even after a solution is in-hand (I consider myself among them, for example). There is value in non-hard-questioners reading this explanation - even if it begins directed at hard-questioners - as the solution makes clear that both sides are making mistakes. We will see that the solution to the hard problem takes advantage of something many non-hard-questioners will have implicitly assumed, but they failed to take that assumption and its implications seriously enough.

With the hard problem clear, there are still some fundamental premises that need to be established. These could be argued at length, but are justified here only briefly:

1. A certain interpretation of Physicalism is true, whereby we refuse to resort to "magic". This just means that we care about reductionism, and believe in the in-principal explainability of all things. Reality is not confused about itself; uncertainty exists in the mind; the map vs the territory, etc. Even if there turns out to exist things like "souls", "heaven", or what we might be tempted to call "immaterial", those previously hypothetical "supernatural" phenomena will just be subsumed into an understanding of what is natural.
    a. We must have some access to what we are talking about when we say the word "consciousness". There is some physical process that is causing us to talk about it, and thus there is some explanation but no guarantee we will ever obtain or understand it. There cannot be "zombies".

   b. Any explanation we put forward has to conform - in some way - with all the empirical findings of both neuroscience and our everyday experience.

2. Consciousness can't be an illusion (given certain definitions of illusion - more on this later in the section on traditional theories, but don't go reading that yet), and in fact is the only thing we know exists for certain. We can be entirely wrong about everything we have ever experienced - via for example living in a simulation, being a brain in a vat, hallucinating, or any other means - but the fact that there is experience itself is impossible to question.

One last thing worth mentioning: this is a conceptual problem - by which I mean this is entirely about untangling a historical, circumstantial, referential confusion - rather than strictly needing to wait for yet more breakthroughs in the "real world" of science/engineering. There are at least 3 features of this type of problem that are worth highlighting via the example of an already solved one: "If a tree falls in a forest and no one is around to hear it, does it make a sound?"

1. We have all the necessary facts and pieces of this puzzle; we don't need some fancy new detector/computer/experiment/etc to solve it. Also, the solution to consciousness doesn't answer everything we care about in philosophy of mind / neuroscience.
2. Little is generated or "depends-on" the outcome/output/resolution of the tree example itself, instead the main end result is to simply stop the conversation about "sound". Now, it's misleading/incorrect to say this document's solution should stop conversations that use the word "consciousness", but it allows us to talk entirely without it; to taboo it. By attaining a solution, there isn't a huge gain in immediate/direct capability, so much as a removal of a huge waste of time; the removal of artificial restrictions born of misunderstanding a domain. To stretch the tree analogy further: imagine there are whole teams of scientists that are holding themselves back from studying the tree, since they were unsure about the outcome/meaning of the tree debate. The dissolution of the tree debate doesn't allow anyone to make "even better" arguments for either side, but rather to stop having the conversation altogether, to stop anyone from tripping-up over fake obstacles, and get back to dealing with reality as best we can.

3. Part of the explanation itself involves describing the nature of the confusion, and thus the circumstantial noises/complaints/arguments the debaters are currently making. Lacking that context, the debaters might be unable to recognize the solution when handed to them, where the solution as an isolated statement can sound banal and/or meaningless.

What We Are

Let's begin.

Why is it like something to be something?

Let's first try to find out what we likely are, as the only things we know with certainty are conscious. If we can first determine what we are, we can then **separately** try and establish how that thing is conscious. Then, we can see if the same criteria work for other things; for answering the hard question in full. (worth reiterating: there are 2 distinct stages here: hazarding a hypothesis about what we are, **then** exploring how or why that thing is conscious; don't prematurely pattern-match this document to something you've already heard about and have learned to dismiss)

We are the software program that is currently implemented by the human brain. Why do we think that? Because it is the most capable and simple explanation on offer, given the tight correspondence among these 3 things:

1. We theoretically and practically know that to have an intelligent and interactive agent in this universe, one prerequisite is to have an internal model of the external reality, which is then manipulated via some program.
2. The empirical findings of neuroscience and the external world's effects upon our body suggest the brain is running such a program.
3. The content of our consciousness correlates with brain states.

Point 1 is justified, given we don't have room in this universe for some alternative solution like a GLUT (Giant Lookup Table). A program can execute without any knowledge of the outside world (excluding implicit "knowledge" of the hardware implementing said program), but for a program to react to the external world it must

gather data about it, then compress and manipulate that data. At the practical level - in attempting to build robots - we have found this to be an engineering necessity.

Point 2 intersects with 1, in that if anything can serve as our hardware for such a program it is clearly the brain (taken to really mean our nervous system at large). Damaging or stimulating anything other than the brain doesn't have nearly the same magnitude or scope of effect on a human's ability to interact with the world as when damaging/stimulating the brain. The rest of the body also doesn't seem to have the right hardware to support computation, while the brain does. Electrical signals are generated and travel from the rest of the body to the brain, and signals are sent out from the brain whose main effects seem to be to directly change some functioning of the body. There is also a great deal of neural activity internal to the brain, which suggests therein resides the calculation and execution of the program, which we wouldn't find if there was some portal to a GLUT, or if the brain was just another pass-through node to some other more active nodes of apparent processing.

Points 1+2 are made from an external/physical perspective, and establish the likely existence of some software program. It may be implicit in the definition of software, but we need to emphasize that our hardware is only important insofar as it implements the software we actually care about. This is true regardless of what one thinks is the true level at which our minds are implemented, for example if you think the brain takes advantage of quantum mechanical effects (unlikely). Even at that level, it is possible to capture those events within a description that abstracts away the algorithms and logical architectures we care about from the substrate; the software from the hardware. If we find the quantum case to be true, we will only have found that this software program is so computationally demanding that the only substrate in this universe powerful enough to implement that program is one that takes advantage of quantum effects, not that the quantum substrate has some magical property independent of its computational power. At a more fundamental level, we know only the software matters due to the universality of computation.

Point 3 is to notice that changes in our supposed hardware change not just our hardware, but our experience of being that hardware. And since hardware only matters due to the software it implements, we can identify/equate ourselves with the software program. Consciousness is the quality/feature of there being "something that it is like to

be something". In our case the **thing** "it is like to be" is the **software program**. If unconvincing, I would revisit the anti-Zombie argument. That article doesn't explicitly argue for this identification with software, but its reasoning can be reused here. If something is running the software program we care about, that then translates to: 1) that thing being capable (given sufficient peripherals/organs) of speaking/reflecting about its consciousness and all varieties of content in its internal experience, and 2) given we are identifying *this* software program, we are not identifying some cheap approximation of its output like a GLUT or a voice recording. So we would have an in-universe, physical explanation for why it is talking about consciousness the same way we do; it is running the same program. To then suppose that this thing wouldn't be conscious - only due to the hardware - would be to suppose the possibility of zombies; something that fulfills all the actually important properties of ourselves, but unlike ourselves isn't conscious. There is no reason to bite this bullet.

(It may sound like I am arguing for Functionalism, which in a certain sense I am, however I address why I refuse this label in the later section on traditional theories. In short, many proponents and critics of Functionalism deliberately apply it to a level of reality too high to capture content we actually care about in our experience. Functionalism (in most of its definitions) can be correct if applied to the right level of reality (and given some extra justification provided herein), but as a historical term is too misleading for me to adopt here.)

All the important content of consciousness we actually care about resides in such a software program. Stated another way, this program is clearly our mind. This isn't begging the question, making a presumptive leap, or trying to smuggle in the word "mind". We aren't yet talking about why or how this mind is conscious. At this stage we are just pointing out that every piece and type of content we actually care about must reside in this software program that is our minds - at a merely functional/physical level. Every emotion, sense impression, thought, and so on must reside in this software program. The question of why that software program is conscious - why there is something that it is like to be that program - remains unanswered.

Let's restate the original question with this new understanding: "Why is it like something to be this software program?"

In dissecting this question, it invokes three perspectives:

- The software program
- What it is like to be the software program; the internal view
- The only implementation we currently have of this program, the brain. This is thus the only external view we have of the program, given neuroscience hasn't yet obtained a full understanding of the brain, which will allow us to "see" the software program itself in full.

How do we reconcile these 3 perspectives?

Dan Dennett vs the Hard-Questioners

Ironically, one way to solve 99% of the hard problem is to ignore one of these perspectives - the internal "what it is like" perspective - and reconcile just the other two. This seems counterproductive, but is remarkable in its ability to get almost all the way there. This isn't to say we ignore content! It is to try to explain consciousness merely by explaining all the external evidence we have of other people's consciousness and content, and to match this with an explanation of a hypothetical software program that contains all the content and action we care about. This is the heterophenomenological approach used by Dan Dennett in his book Consciousness Explained.

Dan Dennett is a non-hard-questioner, and the fundamental communication gap between him and the hard-questioners is that he maps every description of the hard problem into someone defending a cartesian theater, or second transduction, or a kind of elan-vital of the mind. (**TLDR** for those links: he is attacking claims that could take many forms, such as: our inner reality is "another reality" separate from the physical, that there could be things called qualia that are such extra-physical/separate things, that qualia are unexplainable in principle, that there is some redisplay of qualia in an iconic medium for the benefit of the conscious subject (the cartesian theater), etc.)

Those frames of the hard problem are sufficiently different from the hard-questioner's preferred framing that they dismiss any attack on those concepts as being irrelevant to the actual problem they care about. They may even admit that Dennett is effective in refuting those concepts, but still hold that this has no bearing on the actual hard

problem. Then Dennett continues in his own confusion by insisting that every description of the hard problem he hears sounds exactly like a kind of cartesian theater or second transduction, and thus the hard-questioners are simply failing to see this; they are failing to try imagining hard enough just how their problem could actually be removed. Deliciously, I think we will see both sides are partially right.

The real power of Dennett, however, is not in his attacks against the hard problem, but in his positive account of how to approach explaining what he regards as actual problems. In effect, what he does is lay out a description of a hypothetical software program in exquisite detail, such that we gain an intuition for the extent of a mere program's abilities to account for the content of consciousness we care about.

From here on I will take Dennett as a representative of this non-hard-questioner + software program view, without being tied to his particular formulation. For example I consider the predictive processing theories advanced by Anil Seth and others as members of this class. Keeping that in mind, when I refer to a "Fictional Dennett" I mean for him to act as a stand-in for this entire class, instead of always saying what I expect actual Dennett might say.

The conversation at this point looks something like the following:

Fictional Dennett: "Imagine the sensory electrical signal entering the brain. Imagine how it takes up some mode of representation. See how all that remains are further representations and manipulations of those representations by the program running in the brain. Look - here is a hypothetical sub program for this sub problem. Look, can you not imagine how this could be solved? Look at this solution." Then repeat this for every other supposed "easy" problem.

Or, as an actual quote from one of the best Dennett talks available online: "...that's how the brain represents information...and that's all there is to it."

But he misses the most important part! Even after you find how all the things in our mind can have some representation in the software program, there is still the question of why there is consciousness of any of these representations. What goes unstated is WHY having a software program for any one of these things would explain consciousness.

Fictional Dennett: "Look! Look at it! The software program exists! What is the problem? Are you not able to imagine this in enough suggestive detail for it to be convincing? Let me spend a couple hundred pages holding you by the hand so you could visualize this software program and all its capabilities in more detail until you get it."

All the while, the hard-questioner is patiently sitting through all these expositions and thought-experiments surrounding this software program - totally unaffected - and waiting to drop the line: "Why is that software program conscious?"

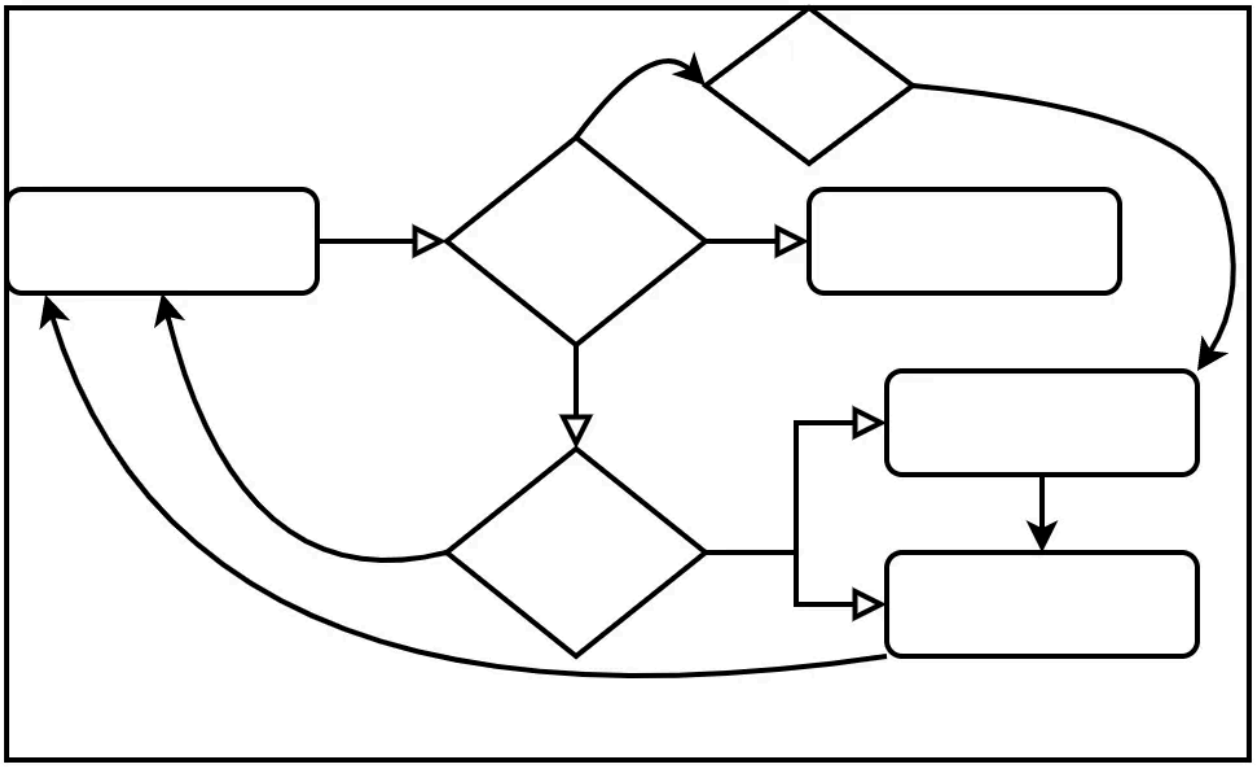
Fictional Dennett: "The software program clearly exists! That is all you need! Let me show you why that is all you need."

But Dennett then repeats the mistake of diving back into a description of the software program. His frame is too functional/physical and doesn't touch the somewhat **ontological** concerns of the hard-questioners. He doesn't deal with the existence of either our subjective experience or the software program directly enough.

The software program can be shown to be **capable** of making a meat puppet do all of the things we witness in ourselves and others - including movement, intelligent speech, and even things like having emotions and internal reflective states (which is conceptually independent of whether there is consciousness of those internal states). But the fact that the software program has these **capabilities** doesn't address the actual hard problem: why is there an "inner reality" that accompanies all this function?

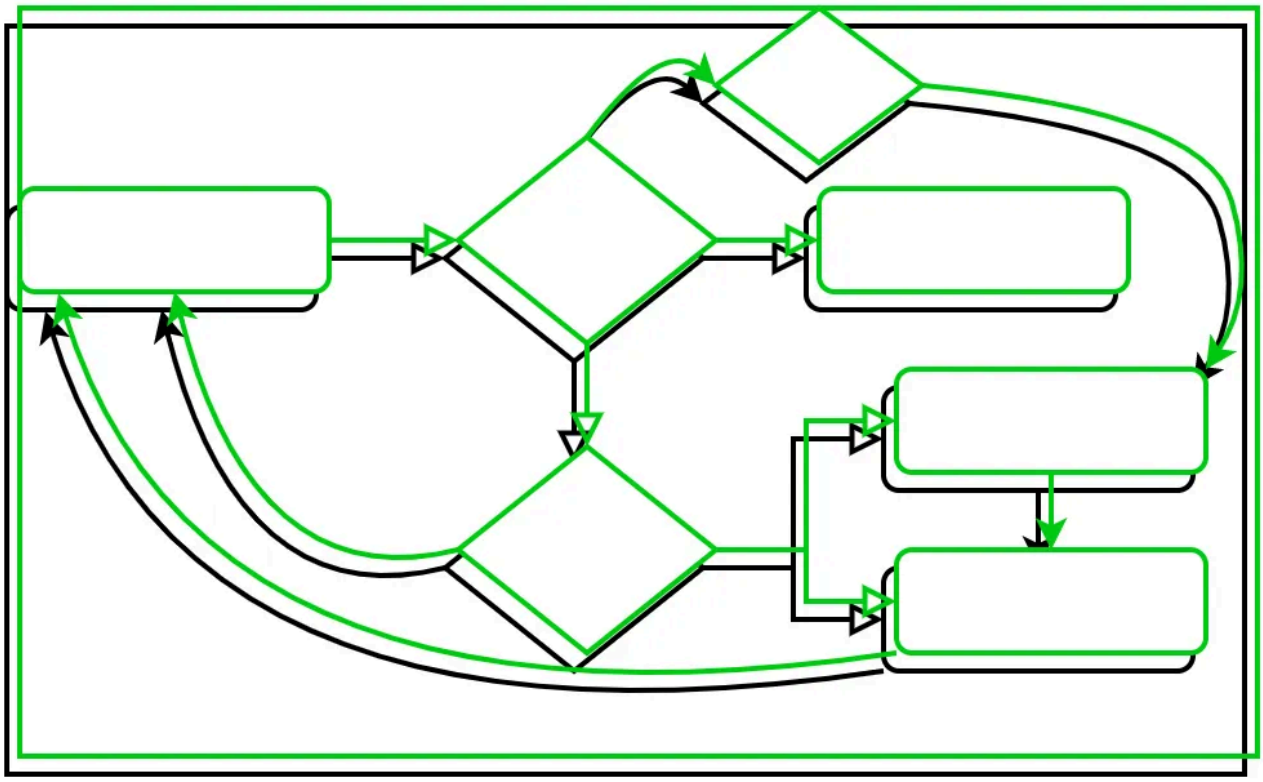
The non-hard-questioner has an intuition that "this is all you need" to explain consciousness, and doesn't know how to deal with the seemingly vacuous opposing intuition of the hard-questioner that "something is left out". Both sides are partially right: there is something further that needs to be explained, but that something is more subtle than both sides think, is less magical than the hard-questioner thinks it will have to be, and is partially taken for granted by the non-hard-questioner. But the non-hard-questioner doesn't appreciate the full heft of what they have taken for granted or what the hard-questioner is actually pointing towards.

Unfortunately, the hard-questioner has placed themselves into a seemingly impossible conundrum. Imagine the hypothetical complete blueprints of our program:



The steelmanned hard-questioner wouldn't immediately dismiss some supposed explanation emerging from within the box of this functional machine. They are humble enough to suppose that perhaps their imagination failed to account for a solution from within the system - like how 'liquidity' can be explained by lower-level molecular motion - but every explanation for consciousness they have ever heard ends up squarely within this functional box. The steelmanned hard-questioner will have the patience to listen, even if within a second of understanding the proposal it has been pattern matched as "merely functional".

To take the hard-questioner at his word, it sounds like he is proposing something orthogonal to the functioning of the program that somehow still informs the program - something outside the diagram that still somehow affects the boxes/arrows. Imagine we changed, or continuously kept changing, the color of the diagram.



There is nothing we could draw within the logic of the diagram that could let it know what the current color is. There is no node/box/arrow that could reference that state. The color is truly independent.

But wait, a steelmanned hard-questioner would be a physicalist, who would reject zombie arguments - so why do they keep making this objection - dubbed “The Tibetan Prayer Wheel”: ?

> “...an objection that is so perennially popular at moments like these that the philosopher Peter Bieri (1990) has dubbed it The Tibetan Prayer Wheel. It just keeps recurring, over and over, no matter what theory has been put forward:
>
> ‘That’s all very well, all those functional details about how the brain does this and that, but I can imagine all that happening in an entity without the occurrence of any real consciousness!’
>
> A good answer to this, but one seldom heard, is: Oh, can you? How do you know? How do you know you’ve imagined ‘all that’ in sufficient detail, and with sufficient

> attention to all the implications? What makes you think your claim is a premise leading to any interesting conclusion?" - Dan Dennett in Consciousness Explained

What happens in the limit of our understanding, when even the hard-questioner's own vocalizations against the functionalist-box can be explained by tracing the causality of the generation of their statements within that very box? At that point we would have to recognize the functionalist-box had won. But if we can already reasonably anticipate that future update, is there anything at all of value to be recovered from the hard-questioner?

The hard-questioner - in that future - would learn *something*. The steelman hard-questioner's model would seemingly have to adapt in some direction in response to that information, but since they are equally capable of imagining this future, they may still try to re-invoke the Tibetan prayer wheel. But suppose there is a reason hard-questioners keep grasping for some orthogonal thing. Is there anything to be gained by taking what they say seriously; giving them the benefit of doubt?

This means hunting for something orthogonal, but something we can still refer to. Seems to be an impossibility. (aside: it is the pressure of this impossibility that has forced some to reach for crazy-sounding theories like panpsychism, which I will address later).

Now I ask for your patience - as this will require several passes to fully grok - but there *is* such an orthogonal property: "Existence". Going back to our functional-box: we can see that a causal node can implicitly and explicitly recognize the existence of the other causal nodes it interacts with. Existence is of course a precondition for anything at all to happen, but that is no obstacle to our ability to *recognize* that precondition. Such recognition can take place within the functional paradigm - so there is a causal chain that interacts with that recognition - but the property we are recognizing is itself outside the causal chain.

To see existence as only a precondition, rather than as something that can be realized at multiple "levels", is a mistake. Existence is a feature of any particular thing. It is not justice, coherence, solidity, consistency, love, fairness, squareness, purpleness, tastiness, or any other similarly abstract concept. What if there was some fundamental puzzle

regarding squares, and I were to provide a solution that depended on the fact that squares have straight sides? Would calling attention - or singling out - this sub-feature of squares as key to the solution be considered a non sequitur? (ex: "Of course the solution had to involve straight sides, this is a square we are talking about after all! The straightness is a precondition"). No! This "singling out" is perfectly valid. In our case we are singling out existence as a concept/property/feature, which is also a property that can be recognized by agents like ourselves. We are unknowingly invoking it when we reflect on what we have access to - our internal model - and call the recognition of that existence our consciousness.

What addresses the hard problem is: the fact that the software program itself actually exists. *This is the answer, but it is sufficiently subtle so as to be easily misunderstood.* We must insist that it is not merely the brain which exists, but the software program that the brain implements exists. The **existence** of the software program **is** the "inner **reality**" that befuddles the hard-questioner. This doesn't require some magical "extra-existence" of the program alone; this existence is just a restatement of what is already there. Not taking that existence seriously enough - whether by imagining the program is merely a useful metaphor, or a useful abstraction, or a human-invented interpretation of the physical events in the brain, or by any other means - is the problem.

Once you have equated "the existence of the software program" with "our consciousness", the full problem is almost ready to melt away. What remains is to realize that there is nothing special - as far as consciousness alone is concerned - with the mere capabilities of the software program. What matters is existence, full stop. So just as the software program exists, so too do rocks, atoms, galaxies, and well... everything. Identifying and equating "existence" with "consciousness" is how you actually solve the hard problem in all its variants. Consciousness is existence, but what exists in our case is the software program that is implemented by our brains. Further exploration of this identity solves all *relevant* problems/confusions regarding consciousness that have ever been raised.

Now hold on! I can feel your derision. Something like: "Good lord this is just panpsychism with extra steps!" But please, let me demonstrate that nothing spooky or supernatural is happening. At some point I had to explicitly state what the answer is, and I'd rather have the answer in mind while I continue to shore-up the arguments that

support it, rather than beating around the bush until the very end. We are not elevating the importance of all atoms to our level. If you like, we are instead lessening the importance of the *very specific* concept that is consciousness; we are demystifying it, by realizing we were using two words for the same thing viewed from two different perspectives. Rather than all of existence suddenly jumping up in value, I would rather you have the intuition that consciousness is relegated to the much more mundane, plentiful, and normal status of existence. We don't lose anything of value, and I promise it all adds up to normality. Suspend your disbelief while I attempt to impart this intuition. Also, this answer will leave many, many problems of mind unsolved - and is supposed to! Consciousness is a very narrow and specific concept which has injected its tendrils of confusion into topics it had no business interfering with, as I hope to show.

Re-ask the question: "Why is it like something to be something?", and see its parts:

- The "like something"; our consciousness
- The "to be" aspect; the existence - in this case of the software program.

What happens when you realize these two bullets are equivalent? The hard problem is turned into a tautology; "Why does something exist if it exists?". Like all tautologies, there is no problem to solve once you realize it is a tautology.

The hard problem is only a conceptual confusion. Did we expect anything else? Confusion exists in the mind, not reality. We made ourselves vulnerable to this confusion the moment our species was capable of distinguishing <our perception of reality> vs reality; the map vs the territory. **The ability to ask why our internal model exists, and to mistakenly view that question as fundamentally separate and different from the question of why anything in the objective external world exists, is the main source of this confusion.**

(Point of discursion: However, if we were to instead morph the hard question into a sentence like: "Why does something that exists exist?" we can be misled into another difficulty altogether; the problem of how that hypothetical thing came to exist at all, or why anything exists at all. That is a conversation about physics, metaphysics, and reality itself - and while interesting, we must recognize it is a separate question/discussion.)

Let's get concrete.

Imagine a human looking at an apple. Photons bounce off the apple, enter the retina, are transformed into electrical signals, which are transformed into further neural firing patterns, which are used as the overall program's representation of that apple to whatever degree it needs to be represented (position, color, shape, movement, its gestalt and association with other objects, association with memories, etc). There is a representation of the apple, and our "consciousness of the apple" is simply our recognition that this representation exists. **It is a restatement of the existence of that part of our internal model.** We are able to reflect upon - and form words about - our consciousness of the apple, because that representation is apparently able to interact with the subset of the program that is capable of internal narration. It is thus a mistake to ask the question of where/when "consciousness" enters into the chain of casualty, as we wouldn't ask the same of "existence". The fact that the representation exists - that an instance of that representation is here in reality and thus available to be interacted with - is all that is needed.

There is this impression that the answer to the hard problem must be something DEEP, and indeed it is, as we have stumbled into the leviathan that is the topic of Existence/Being itself. Until now, we didn't realize that is what we are doing. But once we do, we don't need to solve the entire question/problem of existence/being (if there even is a coherent single question to be asked, since the subject is currently so mysterious). Rather, once we realize the mistake we have made, the simple and mundane meaning of everyday "existence" (ex: "Yup, that apple on the table exists") is entirely sufficient to demolish the confusion around consciousness (given sufficient *non-self-sabotaged* reflection on attempts to impart this intuition, like this document). I view this very similarly to Eliezer's post on truth, which describes how most discussions of truth would be better served by everyone returning to their prior simple understanding of the truth that everyone was using before the nature of truth was brought up. Existence occupies similar territory. There are some real consequences, philosophical and otherwise, that come out of a deeper conversation about existence/metaphysics, but most people never need to bother themselves with them, even when they are in directly adjacent territory (such as consciousness / philosophy of mind).

The Necessity of Perspective

A large obstacle to us accepting this answer - that the existence of the program is identical to our consciousness (from here on this claim is dubbed “EC”) - is that we currently only have two perspectives or views of the program; that of **being** the program, and that of externally viewing the only current implementation of that program; the brain. Because these two perspectives appear vastly different, it begs the question of whether we can really justify this identification; whether there is actually only one thing that is merely viewed from 2 different perspectives. Another version of this confusion is when we ask how the brain “gives rise to” our consciousness, as we are somewhat assuming from the beginning that there are two things - one “giving rise to” the other. But remember, we are identifying with the program, not its implementation substrate. Until we are able to more directly understand and “see” the program itself (attaining the level of understanding where we could actually write the program ourselves) it can appear like we are making some dubious mapping between brain states and our conscious experience. Our only current external view of the program is abstruse and lacking in detail and clarity.

But this dissimilarity between the two perspectives is to be expected! Did we think perspective wouldn't matter? The differing perspectives of **being** the program running on the brain vs observing the brain from the outside must exist. So one cannot point to the necessity of different perspectives as *conclusive* evidence that there are two things; there could be just one “object” that has some mapping to those two different perspectives. What remains is to find our mapping.

Imagine a rock, with a 2k camera and 4k camera recording videos of the rock from two different viewing angles, but at the same time. There is remarkable similarity and simultaneity between the recordings - the two perspectives. But any talk of how one recording could "give rise to" or "be identical to" the other recording is bound for confusion. There is one object that gives rise to the two perspectives. In a certain trivial sense one perspective "is identical to" the other - in that they are recording the same rock - but the perspectives are not literally identical given that one recording is of greater fidelity and has a different viewing angle.

It is true that "what it is like to view the object from the outside'' and "what it is like to be the object'' are two separate concepts; there are 2 different and distinct recordings. But there is only one object.

Now one may object that - unlike the 2 nearly identical recordings of the same rock - the two perspectives of neurons vs subjective experience are so very different as to suggest the perspectives are not "capturing" the same thing. But one can construct a more convoluted recording of the rock, say of its gravitational/magnetic fields, or via echolocation, or even the same recordings but viewed as the 0s and 1s of the video file they are instantiated in. There is still a mapping between the single object and these perspectives, it just becomes more difficult to find (and test!) these mappings. Similarly, the perspective of looking at a brain in the external world with our current technology provides only a very vague impression of the software. Imagine trying to determine whether your computer was running Microsoft Word, but your only available tools are external sensors of some kind, and you don't already have the advantage of a monitor, or understanding Microsoft Word or its architecture!

Finding our own mappings is made yet more difficult by the fact that we **are** one of these "recordings"; we inhabit a certain perspective, that of **being** the object itself. In the rock recording example, we've been imagining some mapping in the physical universe between the physical object of interest and some other physical object. Being the object itself still admits of a mapping, it is just a mapping of identity.

Why does this make things difficult? Because if we were just trying to determine the mappings between external objects, then we could begin our investigation with 3 apparent objects: the hypothesized object itself, perspective 1, and perspective 2. And because these 3 would exist in the external physical universe, we can more easily trace their mappings; we can more easily see the matching and overlapping patterns. We can pick up a DVD and see the pits and lands, or pixels, 0s and 1s, or any other medium of storage.

Our own case is not hopeless, since we have begun to map neural correlates to conscious experience - from simply feeling a blow to the head, to advanced neuroimaging and direct brain stimulation. But we cannot "look" at the perspective of being ourselves through any other lens besides that of simply being that perspective. With the recording

we have the advantage that we can literally pick up the recording medium (ex: DVD), we can see the tiny pits and lands in the disc, we can see how those marks are interpreted by a computer and rendered visible to us with the aid of a computer screen. We can see the recording itself as an object in the world, and this makes the process of determining whether 2 given recordings are actually recordings of the same thing vastly easier. In our case, there is no other "perspective" through which we can view the direct perspective that is ourselves. Our only "perception" of our direct perspective is the perspective itself.

The hard question is misplaced/misphrased in part because it presupposes that what it means “to be something" is a settled and completely understood fact, and then coming out of the blue we have this seperate mystery of why “is it like something to be something". By logical necessity we can only ever be one thing at a time; we are whatever it is we are, at any given time. And the *one thing that we are has a perspective*! What a coincidence! The *only* “experience” we have of **being something** is of being ourselves -- and it turns out that there is **something that it is like** to be ourselves. And the two change in lockstep! What a coincidence!

What can be mistaken as another "phenomenal/magical" realm is just a matter of perspective: to BE the brain state corresponding to apprehending the "blueness of blue" vs the perspective of looking at that brain, or of the state diagram of the program, etc. Want to recover the "magical/phenomenal"? Then BECOME the thing you are concerned with. BE the pattern/representation - inhabit that perspective - and what will exist "for you" will be very different from an outside observer, even though we can zoom-out and see that all of this - including the outside observer and the rest of the universe - also of course exists as part of the same physical universe. There isn't anything magical or mysterious about "perspective" - or invoking some special kind of locus for a subject/agent.

There is always the possibility of multiple perspectives of a single object, so upon finding two perspectives you cannot immediately conclude that there are 2 objects, and appealing to the vast differences between 2 perspectives also isn't a conclusive argument for 2 objects, but it can still be suggestive or count as evidence for such. Determining the validity of a mapping would take the route of normal scientific and probabilistic investigation, which prevents us from making arbitrary mappings between

any 2 things. I can make the hypothesis that the firing of my neuron X is determined by the movement of air particle Y, but in all likelihood any correlation between the two will quickly diverge, and that hypothesis will be discarded.

Similarly, one can propose theories about the mapping between the brain and conscious experience that end up being false, but that doesn't mean the entire project of finding a mapping is doomed. Instead, only that particular interpretation of the brain's activities would be false. What would it take to show that no mapping is possible, rather than any particular mapping theory being incorrect? There is no absolute point at which we would give up, but eventually the failure to generate any true mapping statements or predictions would provide sufficient evidence of the impossibility of mapping brain states, just like any other scientific theory. But we have already found fairly stable brain mappings at various levels of fidelity. From getting a concussion to selectively stimulating tiny sections of brain tissue, we can see these effects on our consciousness and are slowly building our understanding of this mapping; of how the program is implemented.

Neuroscience will Find a Convincing Interpretation

So here's a prediction: as neuroscience continues, these mappings will only ever grow more detailed and accurate, and in time they will nearly completely destroy any motivation for persisting with the hard problem. This is because a major motivator for the hard problem is this impression I've just described; the supposed inability to reconcile the external vs internal views of the brain. But what happens when neuroscience captures and recovers **all** of the content of our consciousness, and renders it viewable through some external medium? We will see that neuroscience alone provides a path to understanding that consciousness is existence.

When hard-questioners operate only within the frame of "consciousness vs <content of consciousness>", most will agree that neuroscience will eventually understand everything physical about the brain, everything about the program, the solution of every easy problem, and the representation of all content. But they remain uncertain or outright against the idea that we would then necessarily be able to "see", prove, or understand any consciousness of that content from this external perspective.

In response to that claim, we must implore them to really imagine in detail what such "mere" neuroscientific progress would actually look like. The hard questioner has no problem granting that eventually we will discover the full contents of consciousness *to the same level of detail as we experience them*. Truly reflecting on the meaning of that prediction should make any hard-questioner nervous.

We will find an interpretation of the substrate - our brain - that renders a version of viewing our conscious experience in full. Take an example experience: that of walking into a grocery store while thinking about a friend, and keeping a shopping list in mind. What happens when neuroscience can directly find all the representations of all the attendant content of that experience? When the visual field can be fed to an external monitor in full detail? When every touch, sound, taste, and smell can be rendered externally "viewable" via the combination of something like a haptic feedback suit, speakers, and either chemical application to - or direct electrical stimulation of - our taste and smell organs? When proprioception and equilibrioception are approximated with the same haptic suit and/or manipulations of our inner ear? When every subvocalization, stray thought, and memories of past sense-impressions and events are rendered viewable via the same approaches? When approximations of emotions are mapped to something like a color wheel, and/or direct-to-text descriptions of the current emotional state? When all of the above content can be *predicted in advance*, down to the level of detail of *the exact wording the subject will use* to describe their experience? Admittedly, once you start getting to more vague or exotic mental content - such as "gut feelings" - it does get harder to imagine how we would render that externally viewable. But to the extent that such content **is there**, we would be able to recover it.

(Caveat / point of discursion 1: Wait a minute, isn't this Dennett's brainstorm machine, which seems to show that we couldn't recover the exact experience (as the hard-questioner would think of it) of certain content? Say, color? This is a distraction which I address in the later section on traditional theories. Don't go read that yet; for now just know that it's possible for one to be mistaken about what form our mental content/representation must take, and how much detail/information *must* be present to produce any given experience.)

(Caveat 2: It would be comparatively easy to cheat in some areas of this "reproduction of experience". For example you could put a subject in an equally complex haptic *recording*

suit, with a camera, microphone, chemical sampler, etc. - that records all incoming events before they even reach the senses. Then "reproduction of experience" merely consists of repeating approximations of those input events. This would effectively be using the mind/brain of the subject for whom the experience is being reproduced to create the experience in full. This just begs the question, as this approach treats the brain as a black box; a simple receiver of sense-impressions. Rather than pointing a camera at the sky and falsely claiming we can thusly reproduce the conscious experience of looking at the sky, my imagined approach would instead only look at internal brain activity - at the software representations of that content - and reproduce an interpretation of that content that recovers all aspects of the experience we identify with "looking at the sky".)

I expect such technology to be wholly convincing to most. **In finding the content, we will have found the consciousness of that content in full.** We will have found our conscious experience. This is a restatement of realizing that consciousness is just existence, but what exists in our case is the program. Consciousness vs content is thus revealed to partially be a false distinction. Most of the time when we talk of there **being** things (ex: "There is an apple on the table.") we don't go out of our way to say: "Hey by the way, regarding that thing that I just said is there, it also exists". The concept of "existence" is still valid, but we take it for granted in most situations.

Me: There is the content. Done.

Hard-questioner: But wait, why is there an inner experience of that content?

Me: That content **is there**. Taboo “experience” and you may unfortunately re-beg the hard question in full, but there is another way you can taboo "experience": what must it mean at a physical level for experience to be taking place? There is the internal model. Done. That internal model itself exists.

Hard-questioner: Yes, and?

Me: Oh, I'm saying that that is what you mean by consciousness. **THERE IS** this inner reality. ***I can point directly to it*** from an external perspective. And what I am ***pointing at*** is the internal model. In identifying that the internal model exists, I am fulfilling all of the same usage and meaning you are referring to when you say you are conscious of

something, you just don't yet realize these terms are perfectly exchangeable. By making this exchange, we demystify "consciousness" as a concept because we now know what we must actually mean when we invoke it.

Both Hard-Questioner and NHQ: Well that still leaves most of what we care about unexplained!

Me: You are right! Consciousness is a very narrow and specific concept that nonetheless has tripped us up. The entirety of the software program still needs to be discovered and understood. All questions of pleasure, pain, value, morality, selfhood, mind design, capability, intelligence, etc. still need to be figured out! But there is no longer any truly deep mystery to our mission of discovery and science moving forward. It is mere engineering all the way down. In fact, the kind of perfect unapproachability evoked by the hard question - for those that fully understood it - should have suggested the possibility that there is something wrong with the question itself. Does an ant or coma patient feel pain? Well, what is pain? And what are its representations in different mind architectures? Asking “is X conscious” is somewhat of a useless question, but questions like “is X conscious of Y” are valid, (even if in need of context-dependent refinement). What remains is to find the Y in question.

The software program is more than a "mere" abstraction. Look at what that abstraction grants us! From the outside we will recover all the content of consciousness, and from the inside we can directly see the reality of that abstraction.

What would we make of the claim that "it is impossible to determine from the patterns of 0's and 1's running through the circuits of a computer whether that computer is currently running Microsoft Word"? We would rightly dismiss this as ridiculous. We have the benefit of already having the right interpretation mechanism built - i.e. the peripherals and UI for msWord - but even if we had never heard of msWord and we are prevented from ever building the right peripherals - such as a monitor, keyboard, or mouse - we could in principle still discover the right interpretation of that circuit activity to "find" msWord out of all the noise. Saying we can't get from the firing of neurons to our conscious experience is merely a statement of our current ignorance regarding the architecture of the software program that is running on our brain.

In the case of msWord, to render that software program useful to its intended user, there must be a monitor and peripheral input devices. But in the case of the software program that is ourselves, the "user" for which the program needs to be "rendered" useful is the program itself. In the case of msWord, we could in principle see the bit-stream and reconstruct all its activity - including what it would look like within its intended "render frame"; what its UI would look like on a monitor. What is the "intended" "render frame" of the software program that is ourselves? There is no external view other than the output of our language, self report, and behavior - but the richness of our conscious experience is just those aspects of the program that need to be rendered intelligible to some subsystem within itself that appears capable of such narration. I say "appears" because it doesn't really matter - for this discussion - what the actual architecture of this program vs its subprograms is, or whether we are accurately capturing any such sub-program with words like "self", “narrative self”, "personhood", "locus of experience", or any other term. Such subprograms are likely useful for reasons of mere evolution and emergent mind engineering, which we aren't concerned with here.

However it seems like we are leaving much unexplained by hand-waving this class of problem away, which in full generality I dub the “problem of access”.

The Problem of Access and Panpsychism

Why are we aware of only what we are currently aware of? Why are we not aware of certain parts of our body, such as our kidneys? Or of aspects of the software program we know must exist, i.e. the subconscious processes that lead to all our action and thought? Why are we not “directly” conscious of other things beyond our body? If consciousness is just existence, why don’t we have this kind of “direct access” to other things that surely exist? Since we can think of the “render frame” as the domain of content of which we are conscious, then if we leave unexplained why our frame has the boundaries that it does, are we not leaving the hard question unanswered?

In answering these, it's useful to first address panpsychism - which suffers from its own version of this problem of access. Panpsychism is an umbrella term for any theory which posits that everything is to some degree conscious, with varying definitions of “everything”. EC is adjacent to certain definitions of panpsychism. Panpsychism alone isn’t enough, as it treats consciousness as either a black box or mysterious extra

property, only that property is supposed to belong to everything. This does solve some problems - enough to make it an attractive option - but not all. Equating consciousness with existence allows us to dissolve every *relevant* conceptual problem and thought experiment that has ever been raised regarding consciousness - including those of panpsychism - because we are no longer poking at a black box.

The largest problem of panpsychism is the combination problem, which is essentially: if everything is conscious, how do "little conscious agents" like quarks combine into higher-order conscious agents such as atoms or ourselves? If consciousness is treated as a black box, it is indeed confusing and impossible to answer how smashing together two little black boxes gets you… a bigger black box? The combination problem is one instantiation of the access problem, in that we can carve out one subset of reality that we believe is conscious (some subcomponent(s) of our software program), but other subsets are possible, and these sets intersect. Is there something that it's like to be my body + the chair, or each individual neuron, or X number of neurons? The choice of subset itself is arbitrary, which seems distasteful until we remember what consciousness is.

Consciousness is just existence, and just as I can say I exist, so too can me+theChair exist. But by stating the realization that me+theChair exists, I am not "calling into being" some new magical agent that presides over me+theChair. We are capable of referring to the existence of subsets of reality without creating new agents or ontological properties - it is just a conceptual and referential convenience to be able to refer to slices of reality. It is a feature of the map, not the territory, that we can say - "Hey, this apple exists" - and to make that statement isn't to call *just that apple* into existence. For us there will only ever be one physical reality, and we are capable of referring to arbitrary subsets of that reality.

Atoms exist, those 5 atoms over there exist, the galaxy exists, my body exists, and the part of the software program that is writing to you now exists.

But what about the part of the software program that *isn't* writing to you right now? It exists, but what would you expect to see differently - *what would you predict would happen* - according to EC? Does anything change? Same for me+theChair; do you expect me+theChair to start speaking to you somehow? How could it do so? Thus any claim

that everything is conscious - at every level of organization and grouping - isn't invalidated by the inability of a system to self report.

The remaining obstacle invoked by the combination problem is a kind of anthropomorphism of experience. In one form, a skeptic trying to make sense of panpsychism will mistakenly try to understand how subsets like atoms could have aspects of consciousness that we are familiar with. A group of 5 atoms isn't capable of having memories, intentions, diverse sensory data, a locus of experience, a subject/self, etc. - these are only contingent facts of the usual contents of human experience, which panpsychism doesn't dictate atoms must have.

The other form of this anthropomorphism is a simple inability to imagine just how minimal experience can get. What would be left for a mere atom to experience? Bridging this gap requires either an active imagination or a willingness to meditate or take psychedelics. Imagine having only a single piece of content within consciousness, say of a single color. No shapes, sounds, touches, body-impressions, thoughts, or memories - just a single unchanging color. The "lights are on", but the only thing "illuminated" is this color; there is consciousness of just this very minimal content. Beyond being able to hypothetically imagine this state, I and others have had this exact experience while under the influence of psychedelics. It's not even that all I could *see* is the color - rather, all that *I was* (at the level of experience) was the color; there was no "I" or me in that moment, there was just the color. Having had such experiences, one realizes just how simple the content of consciousness can get. Simplicity of experience is in fact orthogonal to there being experience; it's not like below a certain level of simplicity everything just winks out. Rather, if there is some content - no matter how minimal - then consciousness is just filled with whatever that thing is.

To summarize, consciousness (existence) does not necessarily entail

- the ability to self report
- a subject/locus of consciousness
- anthropomorphic content of consciousness (things that usually exist in our program, but don't have to exist elsewhere)

I - the subprogram/subset of the entire program running on this brain that is writing to you now - have precisely as much access as I have. I don't currently have access to directly experience my kidneys, and that is a mere apparent fact of neuroscience or neural engineering, not a deep or inherent mystery. It is apparently the case that the architecture of this program doesn't necessitate that <the thing writing to you right now> should have reflective access to its body's kidneys.

To continue to try to imagine some ghost of spooky consciousness presiding over arbitrary subsets - me+myKidneys, me+<the apparently unconscious processing of the rest of the software program>, me+theChair, those 5 atoms, etc. - is a mistake. There is just existence. In fact, just one existence - from which we are capable of referring to arbitrary subsets. One subset we can refer to is the internal model that is ourselves; that we identify as being our conscious content.

Even to ask a question like "What is it like to be an atom?" is somewhat misleading and mistaken. It invokes an image of there being an external vs internal perspective, when really there is just the existence of the atom. Done. To imagine some separate realm or thing that is the experience of the atom - that is *in any way* different from a restatement of the existence of the atom - is a mistake. In our case, *we* have an internal vs external distinction only as a consequence of our ignorance regarding our brain and the software program that we are. If we had the full blueprints and understanding of the software program - *if we could directly "see" the program itself* - that would entail finding all of the content we care about. **We would have found yet another thing that simply exists**, just like apples, or galaxies, or atoms - to the same extent that we can know facts or details or descriptions of those things that we normally regard as existing. We have nearly complete information about what **it is to be** an atom - and it is only incomplete insofar as our understanding of this universe's physics is incomplete; there is not some separate realm of the experience of the atom that needs to be mapped.

Humans have some internal model that still needs to be mapped, but that is just another thing that must exist. *We* (some subset of the program) have apparent boundaries to our content - our "render frame" - as a simple consequence of our apparent engineering. Those parts/processes in our unconscious mind that lead to all our thoughts and actions - of which *we* are unaware - do exist and take up some mode of representation that is manipulated by the overall program, but *we* apparently don't have access to those

representations. Access in our case is the ability to narrate / reflect upon those representations. Those hidden representations are causes of our thoughts and actions, but it is not necessary that *we* should have narrative access to them.

Just as I can program a little AI agent in a video game that only has precisely that level of "access" to the larger world of the entire video game (or the implementing hardware) that I deem useful for its operation in the video game world - so too do *we* (some subset of the program) only have precisely as much access as *we* apparently have. For the access and content that *we* do have, when *we* wonder why that content **is there**, we are really just asking why that content **exists**. As the content changes - or even if *we* intentionally shift our attention to something specific, say for example the sensations in our nose - it is just a subprogram changing what is intelligible or present to itself. Imagine the analogy of an ereader. At any one point there is only a subset of pages open on the screen. The ereader itself could be programmed to turn its own pages without outside input. The currently available contents that are being accessed are the currently open pages. To ask "Well, why is only the current page open?" is not some deep mystery, and we would answer this question by describing the contingent engineering of the ereader program.

Stop thinking in terms of experience or consciousness - think in terms of existence. In this way, the mental image of an atom having some cordoned-off area that is just its experience - the perspective of being just that atom - is nonsensical. We can, for referential convenience, use our map to refer to just the atom - but in doing so we aren't saying that there is something *in-reality* to actually being *just* the atom.

(Aside: I am not re-invoking the flawed concept of a cartesian theater by referring to any such subset of the program; the subset/narrative-layer/whatever isn't another little cartesian agent. If it were possible for a cartesian agent to exist, its imagined purpose is to be the one "doing the experiencing". But after reducing experience to existence, there's nothing to be done to "produce" experience. There is experience. Things exist. Which only leaves "mere" capabilities like a narrative-layer, loci, selfhood, etc to be explained via science/engineering.)

The idea of "pre experiential" vs "experiential" vs "post experiential" brain activity thus somewhat dissolves under EC theory: there is just reality, and some subset of that reality

"makes it" to the point where <the thing talking to you> has access. We only have the access that we have. Regarding, for example, the brain's compensations for head/eye motions to produce the more stable image we see: one flawed way of thinking about it is that those compensations are pre-experiential. But under EC, there is just this chain of causal reality, and the representations we have access to have already been transformed from prior representations that could be interpreted as a more shaky image. Perhaps there is some actual aspect of our narrational system that does indeed have brief access to such shaky representations that are earlier in the causal chain, but that contact/memory is apparently overridden such that all our reports are of a comparatively stable world. This isn't to say that the narrative access layer has the sole privilege of having consciousness. The same property - existence - is present with the shaky representations. The stable representations just make up the narrative-layer/subagent's usual content and are thus what we more readily identify with. But the stable representations don't gain any magic powers by being so selected by our narrative process (suddenly shining in our "inner reality", or suddenly being "actually phenomenally felt"). They get reported as such, but only as a consequence of the fact that *that is all this narrative program has access to; those things are the only things that exist for that program*.

At the risk of annoying via repetition, let's have one last dialogue:

Hard-Questioner: Why *is there* (my outside emphasis) this mental content?

Me: Um… it *is there*. Please reflect on what that actually means. You are asking me to explain something like "why is that apple there?" without resorting to causal physical explanations of the history of that physical apple. Which leaves me only with the answer that "the apple is there". This is thus the same question as: "why does something exist if it exists?" - which again is a tautology that answers itself.

Hard-Questioner: No no I mean of course the "thing" that you are saying *is there* is all the neural firing patterns which we both agree are causing the functioning and the output of this body - all its speech acts and intelligence and so on - but why is there this attendant mental content with all that firing?

Me: Would you ask the same type of question about these two elements?:

- the 0's and 1's currently running on a computer
- the program currently running on a computer ?

What you are doing is the equivalent of asking "Ok but why is the UI for msWord currently displayed?" and/or "why does the UI have certain quality/feature X?". But at the same time you don't want to hear "merely" functional answers to this question like "well the 0's and 1's are organized/patterned in a way so as to.....(etc)". You would say something like: "No no I know how the 0's and 1's physically got there, and I know how in aggregate their overall pattern can be interpreted in such a way so as to yield msWord, but there is still a seperate and orthogonal question of why msWord *is there*!" Which of course you would never ask, because in the case of msWord, the apparent mystery completely disappears. After we have explained how msWord "got there", there is no further mystery to be explained.

In shifting our attention back to our own case, we ARE some subset of the program, which makes it harder to accept this answer. Again imagine msWord running on a computer, but with all output (display, audio, etc) removed. MsWord is actually, really, still running. Just because we've removed the human-readable display doesn't mean we can't make truth-statements about what is "on" the "virtual display" of msWord at that moment. All the same data structures and processes are still there and still running.

To shift this thought experiment closer to the human case: what would we make of an alteration of this setup, whereby msWord is allowed to give audio output in the form of english descriptions of its virtual display? What happens when our insatiable critic re-asks the question: "Would the virtual display **really be there**?" How do we answer this question? What kind of answer would our critic actually accept?

In the case of humans, we know there is no iconic medium in the brain (red visual input isn't actually red in the brain, sounds aren't reproduced in the brain as actual sounds reverberating in our skulls, etc); there is no re-display for the benefit of a cartesian theater (no second transduction). There are just representations of these contents, which this part of the program (myself) has access to.

But we can close the loop; we know that there is a reality to *our* "virtual display"! That is our "inner reality"! That is our consciousness! We can shout and plead with others: "I'm

having a conscious experience of staring at this red cup right now! I'm really having an experience of looking at a red cup!" Meanwhile, the audio-only version of msWord is shouting at us: "The cursor is currently hovering over the minimize button! It really is!" Our temptation is to dismiss the pleadings of msWord as merely logically consistent "as-if" descriptions; descriptions that are consistent with msWord behaving "as-though" there was an "actual" display it is only just talking about.

The same temptation is immediately rejected in our case. *We* know we aren't "just" composing a verbalization of something that isn't there. What is the primary difference, aside from this felt experience? **We are what we are; we aren't msWord**. We are in a position to verify our verbalizations. But to be in the same "position" with respect to msWord, we would have to **become**, to **be**, msWord. Again, this isn't to say that all logically consistent verbal descriptions of some hypothetical world actually mean there is some actual instantiation of that model/representation/program. The precise content/representation and program actually matters; we are not a GLUT.

Our software program is more than a "mere" abstraction. Look at what that abstraction grants us! From the outside we will recover all the content of consciousness, and from the inside we can directly see the reality of that abstraction. Look again upon the original confusion: "Why is it like something to be something?", and see that our ability to ask why our internal model exists - and to mistakenly view that question as fundamentally separate and different from the question of why anything in the objective external world exists - is the main source of this confusion.

Taking Stock

In the end, there may be two equivalent ways of talking about consciousness:

1. There is phenomenal experience - and both we and everything else has it. Consciousness is special, and upon investigation (once we figure out that what we actually mean by "consciousness" is existence - in our case the existence of our model/program), we realize that this specialness is shared by everything that exists (but not necessarily value/pain/pleasure/good/evil/etc - which is just particular content).

2. OR, you could say there is no phenomenality - and that fact is shared by both us and everything else. What actually makes consciousness so special (while being careful to divorce consciousness from particular contents, i.e. value/pain/pleasure/good/evil/etc)? What do you think is special? Why do you think it is special? You only have exactly as much access as you have, and you deem what you have special! If you were to have new experiences (get more information) you would also allow that within the umbrella of this specialness - that "inner light/world" of subjective phenomena/qualia! Isn't that curious? What if something *outside* your access/purview also "glowed" with specialness? *How would you find out about that? You don't have access!* From Dennett: “The *sort* of difference that people imagine there to be between any machine and any human experiencer (recall a wine-tasting machine) is one I am firmly denying: There is no such sort of difference. There just seems to be.”

Hard questioners begin by being ensnared by a conceptual confusion. But it would be inaccurate to say that traditional non-hard-questioners really understand the nature of that confusion and how to solve it. And in light of EC - or whichever version of this solution solidifies in the zeitgeist - I still fear they will claim something to the effect of "Oh, that is what was tripping you up? Of course that is implicit and assumed by our theories! This is basically a version of what we've been trying to say all along". In other words, they will try claiming their prior mere *dismissals* of the hard problem can be rounded-off to - or equivocated with - this explanation that actually *dissolves* the hard problem.

The real proof that the non-hard-questioners weren't really understanding the actual solution is that they failed to see the implication that everything that exists is conscious. That consciousness and existence are the same thing - just viewed from different perspectives. It is true that some non-hard-questioners correctly - if implicitly - assumed that the software program exists. But they failed to grant enough *gravity* to that existence; to take that existence seriously enough. As I’ve said, this doesn’t mean that everything has a mind, or pain/pleasure, agency, value, etc. Questions like "how could we tell a conscious rock from an unconscious rock?" are revealed to be nonsensical, both because this theory says everything is conscious, and more fundamentally because according to any theory we wouldn't expect a rock to "behave" differently if we discovered it was conscious.

Also, not all intelligent minds in design space would be inclined towards our fascination with this hard "problem" of consciousness. Not all minds will be confused about the same things. A random alien/robot/AI might not take a second glance at the fact of its own awareness, and/or how it can be aware of its awareness, or awareness of the map vs territory, etc. In our own way we somewhat already know this, since we have non-hard-questioners in our midst - whose native intuitions can't begin to make sense of the endless ramblings of the hard-questioners ("why does the apple exist if it exists!").

Finally, for everyone to march forward without a backwards glance at the hard problem and its solution is a grave mistake, as this solution itself suggests ways to solve the other deep problem of "why there is something rather than nothing" (aside: there should be some shorthand name for this extremely common question, and it's somewhat ridiculous that there doesn't appear to be one - as even the wikipedia article just reproduces the entire question sentence). Admittedly, the solution to this question will appear to most to be even more unacceptably spooky / magical than EC, although I will maintain it is actually perfectly reductive and rational. Binding together the answers to these two deep questions grants them a strength and coherence they would lack individually, but I know that doing so here will only needlessly increase the attack-surface of EC. And really, this explanation of the hard problem can stand on its own, independent of any metaphysical musings. Whatever can be said about other kinds of metaphysical existences/realities, **this** software program clearly exists and is implemented in this mundane physical universe, and we are that program.

Addressing Traditional and Other Theories

This document is already too long to dissolve other common thought experiments regarding consciousness - or to further relate EC to other theories - so here is a separate document which attempts to do both. To be completely honest, my original intention/motivation to dig through every historical theory quickly died out, so some entries are sparse. But take a look.

From my perspective, EC has reached the level of a nearly unassailably deep intuition - such that I view failures of this document as only failures of my ability/motivation to impart this intuition, or calculated redactions to keep it short. That being said, in attempting to seriously ask myself the question "What could disprove EC?", only one

possibility suggests itself to me: if our understanding of the software program never coheres for **conclusive** reasons, and/or we fail to find a program at all, or we fail to "recover" some very blatant conscious content out of the program. We seem somewhat past this point of possibility, since we already have some rudimentary understanding of the program, such that any further mystery would suggest to us a failing of our own understanding rather than that our pattern of neuronal firing is pure noise, or the firing's true cause is coming from some unexplained magical source. Really, this is the only realistic counter-evidence I can think of. Once we've found the content, there is nothing else we need to find, and all our consciousness-relevant questions will be answered.

1 Like

3 Comments

Write a comment...

anhinga anhinga's Substack May 5 · *edited May 5*

The main weakness of this text seems to be that it does not address qualia (the nature of qualia and their "subjective texture").

Of course, this is the weakness shared by 99%+ of the "theories of consciousness".

99%+ of the "theory of consciousnes" only speak of whether something is or is not conscious. But that's not the most interesting question, the most interesting question is **how** does it feel to be that something, not whether it feels like anything at all.

The "hard problem of qualia" is the hard core of the hard problem of consciousness. For example, we understand why the space of colors is three-dimensional, but we don't understand why any particular color mixture has this particular subjective color for me. And we don't understand why a particular strange electronic sound feels this particular way for me, and so on...

Almost all attempts to solve the hard problem of consciousness ignore the "hard problem of qualia", and I think this is the reason why they don't generally seem to lead to much

progress in our understanding.

If we understand qualia (quale is an element of subjective feeling which is felt by some entity, we want to understand what is the nature of those elements and what they are), we would probably be able to figure out the rest (why qualia tend to group into single consciousnesses, and so on).

LIKE REPLY SHARE

The Ancient Geek RationalityDoneRight Jan 24

It seems that "existence is consciousness" is doing all the lifting here. If we weren't software programs, we would still exist and be conscious.

And I am not clear about what motivates the idea that we are software. Software/Code doesn't really have an ontological existence separate from matter/hardware, it's more of a stance or abstraction. With the hard problem clear, there are still some fundamental premises that need to be established. These could be argued at length, but are justified here only briefly:

A certain interpretation of Physicalism is true, whereby we refuse to resort to “magic”. This just means that we care about reductionism, and believe in the in-principal

You have not put forward a reductive explanation of consciousness, because there is no physical reason that every existing thing should just be conscious.

What is hard about the hard problem is the requirement to explain consciousness, particularly conscious experience, in terms of a physical ontology. Its the combination of the two that makes it hard. Which is to say that the problem can be sidestepped by either denying consciousness, or adopting a non-physicalist ontology.

Examples of non-physical ontologies include dualism, panpsychism and idealism . These are not faced with the Hard Problem, as such, because they are able to say that subjective, qualia, just are what they are, without facing any need to offer a reductive explanation of them. But they have problems of their own, mainly that physicalism is so successful in other areas.

LIKE REPLY SHARE

1 more comment...