The Eliminativist Approach to Consciousness

Brian Tomasik
Foundational Research Institute
brian.tomasik@foundational-research.org

Abstract

This essay explains my version of an eliminativist approach to understanding consciousness. It suggests that we stop thinking in terms of "conscious" and "unconscious" and instead look at physical systems for what they are and what they can do. This perspective dissolves some biases in our usual perspective and shows us that the world is not composed of conscious minds moving through unconscious matter, but rather, the world is a unified whole, with some sub-processes being more fancy and self-reflective than others. I think eliminativism should be combined with more intuitive understandings of consciousness to ensure that its moral applications stay on the right track.

Contents

1 Introduction

> "[Qualia] have seemed to be very significant properties to some theorists because they have seemed to provide an insurmountable and unavoidable stumbling block to functionalism, or more broadly, to materialism, or more broadly still, to any purely 'third-person' objective viewpoint or approach to the world (Nagel, 1986). Theorists of the contrary persuasion have patiently and ingeniously knocked down all the arguments, and said most of the right things, but they have made a tactical error, I am claiming, of saying in one way or another: 'We theorists can handle those qualia you talk about just fine; we will show that you are just slightly in error about the nature of qualia.' What they ought to have said is: 'What qualia?'"
> Dennett (1988)

My views on consciousness are sometimes confusing to readers, so I try to explain them in different ways using different language. I myself also try to imagine the situation from different angles. Three main perspectives that I've advanced are

- reductionism (mainly with a functionalist flavor): consciousness *is* certain algorithms that physical processes perform to varying degrees
- eliminativism: the focus of the current essay
- panpsychism: consciousness is intrinsic to computation, having different flavors for different computational systems.

Pete Mandik has a nice video explaining the distinction between reductionism and eliminativism.

That said, I think all three of these approaches are substantively the same, and they differ mainly in the words they use and the imagery they evoke. These differences may have practical consequences insofar our moral intuitions depend on how we think about consciousness, but what the viewpoints actually say about the world is identical in each case. To make this clear, consider the classic analogy of élan vital. We can pursue any of the following options with it:

- reduce élan vital by saying that it is the properties that define life
- eliminate *élan vital* as an imprecise concept
- adopt a kind of "pan-vitalism" (or hylozoism) theory according to which everything in the universe has traces of life – after all, even atoms show "behaviors" in response to stimuli, move toward equilibrium, etc.

A similar situation obtains with respect to consciousness.

In "Consciousness and its Place in Nature", David Chalmers recognizes that functionalist-style reductionism and eliminativism are ultimately the same (Chalmers, 2003):

> Type-A materialism sometimes takes the form of eliminativism, holding that consciousness does not exist, and that there are no phenomenal truths. It sometimes takes the form of analytic functionalism or logical behaviorism, holding that consciousness exists, where the concept of "consciousness" is defined in wholly functional or behavioral terms (e.g., where to

> be conscious might be to have certain sorts of access to information, and/or certain sorts of dispositions to make verbal reports). For our purposes, the difference between these two views can be seen as terminological.

Chalmers classifies panpsychism as a Type-F monist view, but I think that functionalist panpsychism is a poetic way of expressing a type-A materialism. That which panpsychism says is fundamental to computation is not a *concrete* thing that could conceivably not be present but is more a way of describing how the rhythms of physics (necessarily) seem to us.

2 Motivating eliminativism

Daniel Dennett is often charged with denying consciousness. Some critics of his book *Consciousness Explained* suggest that its title is missing a word, and it should actually be called *Consciousness Explained Away* (Dennett, 1993). One reply to this allegation is that consciousness is being explained, but it's just not what people thought it was. But I suppose another possible response is to say, "Okay, what if I did explain consciousness away? What would follow from that?" I can imagine an Internet meme of Morpheus saying: "What if I told you that we should get rid of the idea of 'consciousness'?"

Maybe "consciousness" is a word with so much metaphysical baggage and philosophical confusion that it would be best to stop using it. Marvin Minsky thinks so and adds:

> now that we know that the brain has [...] hundreds of different kinds of machinery linked in various ways that we don't understand, it would be a wonderful coincidence if any of the words of common-sense psychology actually described anything that's clearly separate, [...] like "rational" and "emotional" [as] a typical dumbed-down distinction that people use.

"Consciousness" does actually point to some helpful distinctions even given a reductionist world view – just as the contrast between rational and emotional thinking does actually have some grounding in neuroscience (Kahneman, 2011). But "consciousness" can point to a lot of distinctions at once depending on what the speaker has in mind. Maybe we should embrace the eliminativist program and replace "consciousness" with more precise alternative words.

To be clear, my version of eliminativism does *not* say that consciousness doesn't exist. Pace Strawson (2006), it does not "deny the existence of the phenomenon whose existence is more certain than the existence of anything else". Rather, eliminativism says that "consciousness" is not the best concept to use when talking about what minds do. We should replace it with more specific descriptions of how mental operations work and what they accomplish. To again give an analogy with *élan vital*: It's not that life doesn't have a sort of vitality to it; it does. Rather, there are more useful and specific ways to talk about life's vitality than to invoke the élan vital concept. Dennett echoes this in "Quining Qualia":

> Everything real has properties, and since I don't deny the reality of conscious experience, I grant that conscious experience has properties. I grant moreover that each person's states of consciousness have properties in virtue of which those states have the experiential content that they do. That is to say, whenever someone experiences something as being one way rather than another, this is true in virtue of some property of something happening in them at the time, but these properties are so unlike the properties traditionally imputed to consciousness that it would be grossly

> misleading to call any of them the long-sought qualia.

Keith Frankish:

> I don't think that consciousness is an illusion [...]. The question is what's involved in having those experiences and those sensations. And I think it does [...] beg the question to say that it involves having states with qualia.

Eliminativists remind us that our intuitions about science are not well refined. It's commonly the case that naive notions of physics, biology, and other sciences need to be replaced by more correct, if less intuitive, understandings. Why should it be different in the case of consciousness? People may feel as though they're experts on subjectivity because they are conscious, but they are just one of many conscious minds in the universe. I don't see how this position qualifies one as an expert on consciousness any more than knowing your way around your house qualifies you as an expert on physical space in the galaxy. In any case, the parts of our brains that talk don't even have clear understanding of much of what goes on in our own heads.

In *The Scientific Outlook* (1931), Bertrand Russell wrote:

> Ordinary language is totally unsuited for expressing what physics really asserts, since the words of everyday life are not sufficiently abstract. Only mathematics and mathematical logic can say as little as the physicist means to say.

Eliezer Yudkowsky remembers that his father

> said that physics was math and couldn't even be talked about without math. He talked about how everyone he met tried to invent their own theory of physics and how *annoying* this was.

In a similar way, I claim we can't understand subjectivity without neuroscience (and physics more generally). And while everyone seems to have a pet theory of consciousness (including me plenty of times), these can't substitute for neuroscience.

Our brains are bad at using intuitions about location and properties of an object to describe quantum superpositions. In a similar way, our language of "consciousness" and "qualia" is not suited to precisely describing what happens in the brain.

3 Thinking physically

The language of "consciousness" and "qualia" corresponds to what Philip Robbins and Anthony I. Jack call the "phenomenal stance". In contrast, the eliminativist position corresponds to what Dennett calls the "physical stance".

In breaking our confusions about consciousness, it's helpful to picture the world purely using the physical stance. Stop thinking about raw feels. Think instead about moving atoms, flowing ions, network connectivity, and information transfer. Imagine the world the way neuroscience describes it – because, in fact, this is a relatively precise account of the way the world *is*. If it seems as though everyone should be a zombie, don't worry about that for now.

Compare an insect with a human. Rather than imagining the human as conscious and the insect as not, or even the human as just *more* conscious than the insect, instead picture the two as you would a professional race car versus a child's toy car: as two machines of different sizes, complexities, and abilities that nonetheless share some common features and functionality.

Compare your brain with another part of your nervous system – say the peripheral nerves in your hand. Why is your brain considered "conscious" and your hand not? It's be-

cause only your brain is capable of generating explicit, high-level, and verbalizable thoughts like remarking on its own consciousness. Your hand is also doing neural operations that resemble neural operations in your brain. It's just that the hand's operations don't always get reported via memories and speech, unless they "become famous" within your brain so that they can be thought about and verbalized (Dennett, 2001).

The eliminativist approach encourages us to stop thinking about neural operations as "unconscious" or "conscious". Instead, in humans, think about the pathway along which neural information travels in order to reach your high-level thinking, speech, action, and memory centers. If the information fails to get there, we call it "subliminal" or "unconscious". If it does get there, we call it "conscious" because of the more pronounced effects it can have on other parts of the brain and body. Thus, for humans, we could replace the loaded "conscious" word with something like "globally available"(Baars, 1997).

There are many more details behind what the brain does that we ordinarily think of as consciousness. The best way to get an intuitive sense for them is to learn more neuroscience, perhaps from a popular book. While I became convinced that non-reductive accounts of consciousness could not be right based mainly on philosophical arguments, it was after reading neuroscience that I actually *internalized* the eliminativist world view. The gestalt shift toward eliminativism requires time and reading to sink in.

4 Eliminativist sentience valuation

Picturing systems physically gives us fresh eyes when deciding what we value. When we adopt the common-sense phenomenal stance, we see a world in which discrete minds move about in otherwise unconscious matter. When we adopt the physical stance, we see various kinds of matter interacting with one another. Some of those matter types (e.g., animals and computers) are more dynamic and sophisticated than other types, but there's a fundamental continuity to the picture among all parts of the system. And we can see that while the system can be sliced in various ways to aid in description and conceptualization, it is ultimately a unified whole.

Ethics in this world view involves valuing or disvaluing various operations within the symphony of physics to different degrees. Some philosophers assign value based on the beauty, complexity, or interestingness of the physics that they see. Those who value conscious welfare instead aim to attribute degrees of sentience to different parts of physics and then value them based on the apparent degree of happiness or suffering of those sentient minds. Because it's mistaken to see consciousness as a concrete *thing*, sentience-based valuation, like the other valuation approaches, involves a projection in the mind of the person doing the valuing. But this shouldn't be so troubling, because metaethical anti-realists already knew that ethics as a whole was a projection by the actor onto the world. The eliminativist position just adds that the thing (dis)being valued, consciousness, is itself something of a fiction of the moral agent's invention.

Actually, calling "consciousness" a fiction is too strong. As noted above, "consciousness" refers to real distinctions – e.g., in the case of human-like brains, we may consider global access to and ability to report on information as important components of consciousness. I just mean "fiction" in the same sense as nations, genders, or tables are fictions; they're constructions of the human mind that help conceptually organize physical phenomena.

I should note that making sentience evaluations based on knowledge of physical processes doesn't mean making *superficial* evaluations. A humanoid doll that blinks might look more conscious than a fruit fly, but the 100,000 neu-

rons of the fruit fly encode a vastly more complex and intelligent set of cognitive possibilities than what the doll displays. Judging by objective criteria given *sufficient knowledge* of the underlying systems is less prone to bias than phenomenal-stance attributions.

Moreover, there's a sense in which nothing ethically important would be left out if we eschewed the idea of "consciousness" and only thought in terms of physical processes. In principle, it would still be straightforward to draw ethical distinctions between so-called "conscious" and so-called "unconscious" human minds, because the brain-activity patterns of the two are clearly distinct. We could still hear what people had to say about the intensity of their emotional feelings and use those reports to make judgments. We could watch their brains and see the neural correlates of those reports. We could develop intuitions for what sorts of physical processes lead to attestations of pleasure and pain, and then we could generalize those kinds of algorithms so as to see them in other places. If we in principle had access to all the operations of a mind, there would be no thought or feeling that would go unnoticed. This approach would actually be *more powerful* at locating sentience *even in ourselves* than our subjective feelings are, since the parts of our brains that develop explicit thoughts and decide high-level actions don't have access to most of the neural operations taking place at the lower levels of the brain or other parts of the body, just like they don't have access to the minds of other people or animals. Knowledge of the physical operations taking place in our minds and other minds makes it possible to value processes of which we would have previously been unaware and which may have previously "suffered" in silence.

5 Living in zombieland

Try adopting the physical stance as you go about your day. When you have a particular feeling or become aware of a particular object, think about what kinds of neural operations are occurring in your head as that happens. Contemplate the brain processes that underlie the behaviors of those around you. See yourself as a chunk of physics moving around within a bigger world of physics.

My experience with this exercise is that it soon becomes less weird to adopt a physical stance. It feels more intuitive that, yes, I am an active, intelligent collection of cells whose sharing and processing of signals constitutes the inner life of my mind and allows for a vast repertoire of behaviors. Worries that I should be a zombie vanish, because I can feel what it's like to be physics for what it is. In fact, being physics feels just like it always did when I thought consciousness was somehow special.

In this mood, questions like, "Why do these physical operations *feel like* something?" appear less forceful, because I'm already "at one" with the universe. Yes, the neurons in my brain are doing particular kinds of processing that other clumps of atoms in the world are not, and this explains why these thoughts show up in my head and not in the floor or a beetle outside. But does it matter that I'm having these particular thoughts? Can't other "thoughts" by other parts of physics matter too for being what they are? Isn't it chauvinist to privilege just cognitive operations that are sufficiently complex and of a particular type? Or is extending sympathy to even simple physics a stance that's based more on theoretical elegance and spirituality, when in fact only sufficiently self-reflective minds have "anyone home" who can meaningfully care about his/her own subjective experiences?

These questions are important to debate, but we see that they take place within the eliminativist realm. We can import some

phenomenal-stance intuitions when thinking about what parts of physics we want to think of as "suffering", but we don't trip ourselves up over trying to pigeonhole suffering as being something other than an attribution we make to whirlpools within the ocean of physics.

6 Why this discussion matters

Eliminativism is not universally shared, particularly among philosophers. (It may be more common among neuroscientists and artificial-intelligence researchers?) Sometimes people encourage me to discuss practical questions of sentience with less dependence on my particular philosophical view of consciousness. For example, even if I thought consciousness were a privileged *thing*, I could still argue that basic physics has some small chance of being conscious. This would yield similar practical conclusions as the eliminativist view does.

This is a fair point, but I think attacking the core confusion about consciousness itself is quite important, for the same reason that it's important to break down the confusions behind theism even if you can argue for a lot of the same practical conclusions whether or not theism is true. Viewing consciousness as a definite and special part of the universe is a systematic defect in one's world view, and removing it does have practical consequences. Looking at the universe from a more physical stance has helped me see that even alien artificial intelligences are likely to matter morally, that plants and bacteria have some ethical significance, and that even elementary physical operations might have nonzero (dis)value. In general, Copernican revolutions change our ethical intuitions in possibly profound ways.

7 Does eliminativism eliminate empathy?

A legitimate concern about eliminativism is that it could reduce the intuitive importance or meaningfulness of altruism. If everything is just particles moving in different ways, why should I care? Jack et al. (2013) have found evidence that "there is a physiological constraint on our ability to simultaneously engage two distinct cognitive modes", namely, social and physical reasoning.

But if eliminativism does reduce compassion, it may be because the eliminativist position has not been completely understood. If you still think of consciousness as being something special, then eliminativism sounds like a view that the world doesn't contain that special thing, so it doesn't matter. But what eliminativism really says is that all the specialness you thought was in the world is *still there* and in fact may be more universal than you realized.

Consider a fish suffocating on the deck of a fishing boat. It flops back and forth, apparently in agony. A conventional approach is to say that if this fish is conscious, then it must be aware of its terrible suffering, which is bad and should be avoided. An eliminativist can note how

- aversive sensory inputs are being aggressively broadcast throughout the fish's brain
- the fish's fear system is activated, triggering follow-on changes in its cognition and releasing stress hormones throughout its body
- the fish thinks about ways it might escape from its predicament
- powerful memories of a terrible experience are being created through changes in synaptic strengths and connectivity
- and so on.

When I read this list, I think those brain changes look really bad, and I feel almost as much empathy as when I think about the fish from the common-sense standpoint of having an agonizing subjective experience. If we care about suffering, then we care about what suffering actually is even on closer inspection.

Steven Weinberg said "With or without religion, good people can behave well and bad people can do evil; but for good people to do evil–that takes religion." In a similar way, it's plausible that for good altruists to ignore suffering requires confusion about consciousness. If you hold Descartes's view that animals are non-sentient machines, you can really delude yourself into thinking that the struggling of a dog when vivisected is not conscious and hence doesn't matter. An eliminativist realizes that lots of aversive processing *is really going* on in the dog's head, and so the vivisection must be at least somewhat bad, depending on the negative weight given to those aversive processes. The eliminativist position is thus more cautious in some sense.

That said, I'm describing here mainly my experience with eliminativism, as someone who was already heavily committed to reducing suffering. It remains an empirical question what kinds of effects eliminativism has on average for various populations and depending on the degree to which it's internalized. That said, because I think something like eliminativism on consciousness will become more widely accepted in the future as understanding of neuroscience increases, we may want to figure out how to frame eliminativism in a more altruism-friendly way rather than just sweeping it under the rug.

8 The subjective and objective need each other

The physical stance is more impartial and accurate than the phenomenal stance in accounting for all the mind-like processes that exist in the world. However, the physical stance is also more dispassionate. While the brain of a person being tortured does look physically very distinctive – with lots of activity and long-lasting neural "scars" being created – appreciating its true *awfulness* requires imagining ourselves in its position. Without subjective imagination, a physical-stance approach is liable to give way to aesthetic judgments – valuing more brains that *appear* more interesting, sophisticated, nuanced, or dynamic. Looking for beauty and novelty is a natural temptation when we view physical objects, but it has little to do with ethics. There's a danger that eliminativism gives too much sway to non-empathic judgment criteria.

I think we should try out the eliminativist view as an exercise, to bend our prior prejudices and intuitions. When we unshackle ourselves from the conventional concept of consciousness, how many other ways might there be to reimagine the world! That said, eliminativism doesn't have to be and arguably should not be the *only* way we think about consciousness, just as our slow, utilitarian moral system needn't be the only way we think about ethics (Greene et al., 2001). Rather, we can blend the insights of eliminativism with those of a more common-sense, phenomenal stance – with the aim of achieving a reflective equilibrium that incorporates insights from each.

9 Eliminativism and panpsychism

Eliminativism and panpsychism may seem like polar opposites, but they're actually two sides of the same coin, in a similar way as 0 degrees and 360 degrees on a unit circle point in the same direction. Both maintain that there's nothing distinctive about consciousness that sharply distinguishes it from the rest of the universe. Panpsychism recognizes that all the computations of physics have a fundamental similarity to them, and it considers different computations as different shades of the same basic thing (though the shades may differ quite a bit). Eliminativism rejects talk about "consciousness" in favor of physical descriptions, and once again we can see a fundamental continuity among the diverse flavors of physical processes. Whether it uses the word "consciousness" or not, each perspective points at

the same underlying reality.

That said, it's worth noting that even if we recognize all of physics as fundamentally mental in some sense, it remains a matter of choice how much we care about simple physical operations. We might legitimately decide that only really complex systems like those that emerge in animal brains contain moral significance.

10 Denying consciousness altogether

In the above piece, I tried to insist that eliminativism doesn't deny consciousness per se, only the particular conception of consciousness that some philosophers cling to. In 2015, I'm leaning more toward Minsky's view that it might be most clear to dispense with the "consciousness" word altogether, since it causes so much confusion. Instead, it's more helpful to say: We're not conscious but only think we are.

But how is that possible? "I just know I'm conscious!" But any thoughts you have about your being conscious are fallible. I believe there are bugs in the vast network of computation that produces thoughts like "I'm conscious in a way that generates a hard problem of consciousness." No thought you have is guaranteed to be free from bugs, and it seems more likely – given the basically useless additional complexity of postulating a metaphysically privileged thing called consciousness – to suppose that our attribution of metaphysically privileged consciousness to ourselves is a bug in our cognitive architectures. This is a relatively simple way to escape the whole consciousness conundrum. If it feels weird, that's because the bug in your neural wiring is causing you to reject the idea. Your thoughts exist within the system and can't get outside of it.

Your brain is like a cult leader, and you are its follower. If your brain tells you it's conscious, you believe it. If your brain says there's a special "what-it's-like-ness" to experience beyond mechanical processes, you believe it. You take your cult leader's claims at face value because you can't get outside the cult and see things from any other perspective. Any judgments you make are always subject to revision by the cult leader before being broadcast (Similar analogies help explain the feeling of time's flow, the feeling of free will, etc.). I like how Michael Graziano puts it:

> I believe a major change in our perspective on consciousness may be necessary, a shift from a credulous and egocentric viewpoint to a skeptical and slightly disconcerting one: namely, that we don't actually have inner feelings in the way most of us think we do. [...]
>
> a new perspective on consciousness has emerged in the work of philosophers like Patricia S. Churchland and Daniel C. Dennett. Here's my way of putting it:
>
> How does the brain go beyond processing information to become subjectively aware of information? The answer is: It doesn't. The brain has arrived at a conclusion that is not correct. [...]
>
> You might object that this is a paradox. If awareness is an erroneous impression, isn't it still an impression? And isn't an impression a form of awareness?
>
> But the argument here is that there is no subjective impression; there is only information in a data-processing device. When we look at a red apple, the brain computes information about color. It also computes information about the self and about a (physically incoherent) property of subjective experience. The brain's cognitive machinery accesses that interlinked information and derives several conclusions: There is a self, a me; there is a red thing nearby; there is such a thing as subjective experience; and I have an experience of that red thing. Cognition is captive to those internal models. Such a brain would in-

> escapably conclude it has subjective experience.

So there, I said it: Consciousness doesn't exist. Now let's figure out more precisely what we are pointing at when we seek to reduce conscious suffering.

Often I hear claims that "I'm more certain that I'm conscious than I am about anything else." I disagree. Our perception of being conscious, just like our perception of anything else, is a hypothesis that our brain constructs, based on very complicated processing and lower-level thinking, expressed in terms of a simplified ontology that the brain can make sense of (Gregory, 1980). Anything that you know is the result of complex computation by an information-processing device. But then why privilege some types of visceral, intuitive judgments that your brain makes over other judgements your brain makes? *All* of your knowledge is constructed by the brain's information processing in one way or another. "Knowing that I'm conscious" is not a thought that somehow transcends ordinary brain machinery, nor does it deserve to be made axiomatic in one's ontology.

11 Does eliminativism explain phenomenology?

Focus your attention on the visual image you see of the world in front of you: a rich jumble of colors, shapes, textures, and patterns. How can those not be the philosopher's qualia?

Naively, it looks like eliminativism can't explain these data. But exactly how we characterize the data makes a difference to our theoretical interpretation. If we declare the visual imagery in front of us as something metaphysically special – "mental phenomena" – then eliminativism cannot account for them. But to suppose that the visual scenes we see are phenomena in their own metaphysical category is to beg the question.

An alternate characterization of our visual experiences is that they represent "(data about the external world) + (an explicit or implicit judgment by our brains that we're seeing a rich collection of colors, shapes, etc.)". All we know is the judgments our brains make. If our brains judge that we're seeing rich colors and shapes, then we'll think we are seeing such things. The eliminativist hypothesis, then, just predicts that our brains make these judgments about seeing things-it-calls-qualia when it attends to its processing of visual input.

These judgments needn't be verbal but are often just more basic moments of noticing how something seems. For instance, when I'm going about my day, I typically don't even notice the colors and shapes around me, but if I focus my attention on how they look, I undergo a nonverbal process of feeling like "Wow, there's something it looks like to see what's in front of me!" This feeling is an implicit "judgment" that my brain makes, and it's all that's needed to explain the fact that we feel like we have qualia.

That I typically don't notice "qualia" unless I attend to them bolsters this view. Most of the time, my brain is focused on my own internal thoughts and basically ignores the world it sees. In this case, my brain is just processing visual data "unconsciously". Then, when I focus on some visual input in particular, my brain produces an implicit (and sometimes explicit) judgment that "this thing has distinctive color and texture and shape, and it feels like something to see it". In other words, so-called "conscious experiences" are experiences that our brains judge to be conscious.

Qualia are user illusions. When you click a folder icon on your computer desktop, there's not *actually* a little folder there; your computer just tells you that there's a folder there, when in fact it represents more complex processing in your computer's hard drive. Likewise, when you see a visual scene, there's not *actually* an ontological what-it's-like-ness ex-

perience; it's just that your brain tells you it's having a qualitative experience of the visual scene, when in fact what's happening under the hood is complex brain processing.

Of course, this doesn't mean that only experiences judged to be conscious matter ethically. The judgment process only adds a small level of reflection on what were already complex, substantive computations. The pre-eliminativist view that only "conscious" emotions matter was based on a confused idea that only "conscious" emotions are "real qualia" in a metaphysical sense. Once we cast off this way of thinking, it becomes less plausible that only experiences judged to be conscious have moral weight.

We can describe the same reality at different levels of abstraction, i.e., using different ontological frameworks. As an example of a different ontology, imagine a simple computational agent that moves about in a "grid world" consisting of 16 squares – 16 possible "states" that it can be in. This simple agent knows nothing of the Earth, particle physics, or even humans. It just "knows" (in some very simplistic way) its own little world of state transitions and whatever dynamics drive its behavior. Likewise, we humans are acquainted with a simplified ontology of our own brains – that we have things called "conscious experiences" and that we transition through these experiences. If we were cave people, we would know nothing of neurons, the cerebral cortex, or gamma oscillations (except whatever we observed when eating the brains of animals). We would just have a simplistic, subjective model of our mental lives, which is what people refer to when they talk about knowing that they're conscious. This picture isn't mutually exclusive with a more detailed, physics-based portrait.

References

Baars, B. J. (1997). In the theatre of consciousness. global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, 4(4):292–309.

Chalmers, D. J. (2003). Consciousness and its place in nature. *Blackwell guide to the philosophy of mind*, pages 102–142.

Dennett, D. (2001). Are we explaining consciousness yet? *Cognition*, 79(1):221–237.

Dennett, D. C. (1988). *Quining Qualia. In Consciousness in Contemporary Science, edited by A. Marcel and E. Bisiach*. Oxford: Oxford University Press.

Dennett, D. C. (1993). *Consciousness explained*. Penguin UK.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fmri investigation of emotional engagement in moral judgment. *Science*, 293(5537):2105–2108.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 290(1038):181–197.

Jack, A. I., Dawson, A. J., Begany, K. L., Leckie, R. L., Barry, K. P., Ciccia, A. H., and Snyder, A. Z. (2013). fmri reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage*, 66:385–401.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Russell, B. (1931). *The scientific outlook*. London: George Allen.

Strawson, G. (2006). Realistic monism: Why physicalism entails panpsychism. *Journal of consciousness studies*, 13(10/11):3.