Best of LessWrong: July 2016

1. Zombies Redacted

Best of LessWrong: July 2016

1. Zombies Redacted

Zombies Redacted

I looked at my old post Zombies! Zombies? and it seemed to have some extraneous content. This is a redacted and slightly rewritten version.

Your "zombie", in the philosophical usage of the term, is putatively a being that is exactly like you in every respect—identical behavior, identical speech, identical brain; every atom and quark in *exactly* the same position, moving according to the same causal laws of motion—*except* that your zombie is not conscious.

It is furthermore claimed that if zombies are "conceivable" (a term over which battles are still being fought), then, purely from our knowledge of this "conceivability", we can deduce a priori that consciousness is extra-physical, in a sense to be described below.

See, for example, the SEP entry on Zombies. The "conceivability" of zombies is accepted by a substantial fraction, possibly a majority, of academic philosophers of consciousness.

I once read somewhere, "You are not the one who speaks your thoughts—you are the one who *hears* your thoughts".

If you conceive of "consciousness" as a quiet, passive listening, then the notion of a zombie initially seems easy to imagine. It's someone who lacks the the inner hearer.

Sketching out that intuition in a little more detail:

When you open a refrigerator and find that the orange juice is gone, you think "Darn, I'm out of orange juice." The sound of these words is probably represented in your auditory cortex, as though you'd heard someone else say it.

Why do I think the sound of your inner thoughts is represented in the auditory cortex, as if it were a sound you'd heard? Because, for example, native Chinese speakers can remember longer digit sequences than English-speakers. Chinese digits are all single syllables, and so Chinese speakers can remember around ten digits, versus the famous "seven plus or minus two" for English speakers. There appears to be a loop of repeating sounds back to yourself, a size limit on working memory in the auditory cortex, which is genuinely phoneme-based.

It's not only conceivable in principle, but possibly possible in the next couple of decades, that surgeons will lay a network of neural taps over someone's auditory cortex and read out their internal narrative. Researchers have already tapped the lateral geniculate nucleus of a cat and reconstructed recognizable visual inputs.

So your zombie, being physically identical to you down to the last atom, will open the refrigerator and form auditory cortical patterns for the phonemes "Darn, I'm out of orange juice". On this point, p-zombie advocates agree.

But in the Zombie World, allegedly, there is no one inside to *hear;* the inner listener is missing. The internal narrative is spoken, but unheard. You are not the one who speaks your thoughts, you are the one who hears them.

The Zombie Argument is that if the Zombie World is *possible*—not necessarily physically possible in our universe, just "possible in theory", or "conceivable"—then consciousness must be extra-physical, something over and above mere atoms. Why? Because even if you knew the positions of all the atoms in the universe, you would still have be told, as a *separate* and *additional* fact, that people were conscious—that they had inner listeners—that we were not in the Zombie World.

The technical term for the belief that consciousness is there, but has no effect on the physical world, is *epiphenomenalism.*

Though there are other elements to the zombie argument (I'll deal with them below), I think that the intuition of the inner listener is what first persuades people to zombie-ism. The core notion is simple and easy to access: The lights are on but nobody's home.

Philosophers are appealing to the intuition of the quiet, passive inner listener when they say "Of course the zombie world is imaginable; you know exactly what it would be like."

But just because you don't see a contradiction in the Zombie World at first glance, it doesn't mean that no contradiction is there. Just because you don't see an internal contradiction *yet* within some set of generalizations, is no guarantee that you won't see a contradiction in another 30 seconds. "All odd numbers are prime. Proof: 3 is prime, 5 is prime, 7 is prime..."

So let us ponder the Zombie Argument *a little longer*: Can we think of a counterexample to the assertion "Consciousness has no third-party-detectable causal impact on the world"?

If you close your eyes and concentrate on your inward awareness, you will begin to form thoughts, in your internal narrative, along the lines of "I am aware" and "My awareness is separate from my thoughts" and "I am not the one who speaks my thoughts, but the one who hears them" and "My stream of consciousness is not my consciousness" and "It seems like there is a part of me which I can imagine being eliminated without changing my outward behavior."

You can even say these sentences out loud. In principle, someone with a super-fMRI could probably read the phonemes right out of your auditory cortex; but saying it out loud removes all doubt about whether you have entered the realms of physically visible consequences.

This certainly *seems* like the inner listener is being *caught in the act of listening* by whatever part of you writes the internal narrative, a causally potent neural pattern in your auditory cortex, which can eventually move your lips and flap your tongue.

Imagine that a mysterious race of aliens visit you, and leave you a mysterious black box as a gift. You try poking and prodding the black box, but (as far as you can tell) you never elicit a reaction. You can't make the black box produce gold coins or answer questions. So you conclude that the black box is causally inactive: "For all X, the black box doesn't do X." The black box is an effect, but not a cause; epiphenomenal, without causal potency. In your mind, you test this general hypothesis to see if the generalization is true in some trial cases, and it seems to be true in every one—"Does the black box repair computers? No. Does the black box boil water? No."

But you can *see* the black box; it absorbs light, and weighs heavy in your hand. This, too, is part of the dance of causality. If the black box were wholly outside the causal universe, you wouldn't be able to see it; you would have no way to know it existed; you could not say, "Thanks for the black box." You didn't *think* of this counterexample, when you formulated the general rule: "All X: Black box doesn't do X". But it was there all along.

(Actually, the aliens left you another black box, this one purely epiphenomenal, and you haven't the slightest clue that it's there in your living room. That was their joke.)

If something has no causal effect, you can't know about it. The territory must be causally entangled with the map for the map to correlate with the territory. To 'see' something is to be affected by it. If an allegedly physical thing or property has absolutely no causal impact on the rest of our universe, there's a serious question about whether we can even talk about it, never mind justifiably knowing that it's there.

It is a standard point—which zombie-ist philosophers accept!—that the Zombie World's philosophers, being atom-by-atom identical to our own philosophers, write identical papers about the philosophy of consciousness.

At this point, the Zombie World stops being an intuitive consequence of the idea of an inner listener.

Philosophers writing papers about consciousness would *seem* to be at least one effect of consciousness upon the world. You can argue clever reasons why this is not so, but you have to be clever. You are no longer playing straight to the intuition.

Let's say you'd never heard of the Zombie World and never formed any explicit generalizations about how zombies are supposed to exist. The thought might spontaneously occur to you that, as you stand and watch a beautiful sunset, your awareness of your awareness could be subtracted from you without changing your outward smile. But then ask whether you still think "I am aware of my inner awareness", as a neural pattern in your auditory cortex, and then say it out loud, after the inner awareness has been subtracted. I would not expect the generalization "my inner awareness has no effect on physical things" to still seem intuitive past that point, if you'd never been explicitly indoctrinated with p-zombieism.

Intuitively, we'd suppose that if your inward awareness vanished, your internal narrative would no longer say things like "There is a mysterious listener within me," because the mysterious listener would be gone and you would not be thinking about it. It is usually immediately *after* you focus your awareness on your awareness, that your internal narrative says "I am aware of my awareness"; which suggests that if the first event never happened again, neither would the second.

Once you see the collision between the general rule that consciousness has no effect, to the specific implication that consciousness has no effect on how *you* think about consciousness (in any way that affects your internal narrative that you could choose to say out loud), zombie-ism stops being intuitive. It starts requiring you to postulate strange things.

One strange thing you might postulate is that there's a Zombie Master, a god within the Zombie World who surreptitiously takes control of zombie philosophers and makes them talk and write about consciousness.

Human beings often don't sound all that coherent when talking about consciousness. It might not be that hard to fake. Maybe you could take, as a corpus, one thousand human amateurs trying to discuss consciousness; feed them into a sufficiently powerful but non-reflective machine learning algorithm; and get back discourse about "consciousness" that sounded as sensible as most humans, which is to say, not very.

But this speech about "consciousness" would not be produced *within* the AI. It would be an imitation of someone else talking. You might as well observe that you can make a video recording of David Chalmers (the most formidable advocate of zombieism) and play back the recording. The *cause* that *shaped* the *pattern* of the words in the video recording was Chalmers's consciousness moving his lips; that shaping cause is merely being transmitted through a medium, like sounds passing through air.

A separate, extra Zombie Master is *not* what the philosophical Zombie World postulates. It's asserting that the atoms in the brain are quark-by-quark identical, moving under exactly the same physical laws we know; there's no separate, additional Zombie Master AI Chatbot making the lips move in ways that were copied off the real David Chalmers. The zombie you's lips are talking about consciousness *for the same causal reason* your lips talk about consciousness.

As David Chalmers writes:

> Think of my zombie twin in the universe next door. He talks about conscious experience all the time—in fact, he seems obsessed by it. He spends ridiculous amounts of time hunched over a computer, writing chapter after chapter on the mysteries of consciousness. He often comments on the pleasure he gets from certain sensory qualia, professing a particular love for deep greens and purples. He frequently gets into arguments with zombie materialists, arguing that their position cannot do justice to the realities of conscious experience.
>
> And yet he has no conscious experience at all! In his universe, the materialists are right and he is wrong. Most of his claims about conscious experience are utterly false. But there is certainly a physical or functional explanation of why he makes the claims he makes. After all, his universe is fully law-governed, and no events therein are miraculous, so there must be some explanation of his claims.
>
> …Any explanation of my twin's behavior will equally count as an explanation of my behavior, as the processes inside his body are precisely mirrored by those inside mine. The explanation of his claims obviously does not depend on the existence of consciousness, as there is no consciousness in his world. It follows that the explanation of my claims is also independent of the existence of consciousness.

Chalmers is not arguing *against* zombies; those are his actual beliefs!

> This paradoxical situation is at once delightful and disturbing. It is not obviously fatal to the nonreductive position, but it is at least something that we need to come to grips with...

I would seriously nominate this as the largest bullet ever bitten in the history of time. And that is a backhanded compliment to David Chalmers: A lesser mortal would simply fail to see the implications, or refuse to face them, or rationalize a reason it wasn't so.

Why would anyone bite a bullet that large? Why would anyone postulate unconscious zombies who write papers about consciousness for *exactly the same reason* that our own genuinely conscious philosophers do?

Not because of the first intuition I wrote about, the intuition of the quiet inner listener. That intuition may say that zombies can drive cars or do math or even fall in love, but it doesn't say that zombies write philosophy papers about their quiet inner listeners.

No, the drive to bite *this* bullet comes from an entirely different intuition—the intuition that no matter how many atoms you add up, no matter how many masses and electrical charges interact with each other, they will never *necessarily* produce a subjective sensation of the mysterious redness of red. It may be a fact about our physical universe (Chalmers says) that putting such-and-such atoms into such-and-such a position, *evokes* a sensation of redness; but if so, it is not a *necessary* fact, it is something to be explained above and beyond the motion of the atoms.

But if you consider the second intuition on its own, *without* the intuition of the quiet listener, it is hard to see why irreducibility implies zombie-ism. Maybe there's just a *different kind of stuff,* apart from and additional to atoms, that is *not* causally passive—a soul that actually *does* stuff. A soul that plays a real causal role in why we write about "the mysterious redness of red". Take out the soul, and... well, assuming you just don't fall over in a coma, you certainly won't write any more papers about consciousness!

This is the position taken by Descartes and most other ancient thinkers: The soul is of a different kind, but it *interacts* with the body. Descartes's position is technically known as *substance dualism*—there is a thought-stuff, a mind-stuff, and it is not like atoms; but it is causally potent, interactive, and leaves a visible mark on our universe.

Zombie-ists are *property dualists*—they don't believe in a separate soul; they believe that matter in our universe has additional properties beyond the physical.

"Beyond the physical"? What does that mean? It means the extra properties are there, but they don't influence the motion of the atoms, like the properties of electrical charge or mass. The extra properties are not experimentally detectable by third parties; *you* know you are conscious, from the *inside* of your extra properties, but no scientist can ever directly detect this from outside.

So the additional properties are there, but not causally active. The extra properties do not move atoms around, which is why they can't be detected by third parties.

And that's why we can (allegedly) imagine a universe just like this one, with all the atoms in the same places, but the extra properties missing, such that every atom moves the same as before, but no one is conscious.

The Zombie World might not be *physically* possible, say the zombie-ists—because it is a fact that all the matter in our universe has the extra properties, or obeys the bridging laws that evoke consciousness—but the Zombie World is logically possible: the bridging laws could have been different.

But why, oh why, say that the extra properties are epiphenomenal and undetectable?

We can put this dilemma very sharply: Chalmers believes that there *is* something called consciousness, and this consciousness embodies the true and indescribable substance of the mysterious redness of red. It may be a property beyond mass and

charge, but it's *there*, and it *is* consciousness. Now, having said the above, Chalmers furthermore specifies that this true stuff of consciousness is epiphenomenal, without causal potency—but *why say that?*

Why say that you could subtract this true stuff of consciousness, and leave all the atoms in the same place doing the same things? If that's true, we need some *separate* physical explanation for why Chalmers talks about "the mysterious redness of red". That is, there exists both a mysterious redness of red, which is extra-physical, and *an entirely separate* reason, *within* physics, why Chalmers *talks* about the "mysterious redness of red".

Chalmers does confess that these two things seem like they ought to be related, but why do you need to assert two separate phenomena? Why not just assert one or the other?

Once you've postulated that there is a mysterious redness of red, why not just say that it interacts with your internal narrative and makes you talk about the "mysterious redness of red"?

Isn't Descartes taking the simpler approach, here? The *strictly* simpler approach?

Why postulate an extramaterial soul, *and then* postulate that the soul has no effect on the physical world, *and then* postulate a mysterious unknown *material* process that causes your internal narrative to talk about conscious experience?

Why not postulate the true stuff of consciousness which no amount of mere mechanical atoms can add up to, *and then,* having gone that far already, let this true stuff of consciousness have causal effects like making philosophers talk about consciousness?

I am not endorsing Descartes's view. But at least I can understand where Descartes is coming from. Consciousness seems mysterious, so you postulate a mysterious stuff of consciousness. Fine.

But now the zombie-ists postulate that this mysterious stuff *doesn't do anything*, so you need a *whole new* explanation for why you *say* you're conscious.

That isn't vitalism. That's something so bizarre that vitalists would spit out their coffee. "When fires burn, they release phlogiston. *But* phlogiston doesn't have any experimentally detectable impact on our universe, so you'll have to go looking for a separate explanation of why a fire can melt snow." *What?*

Are property dualists under the impression that if they postulate a new *active* force, something that has a causal impact on physics, they will be sticking their necks out too far?

Me, I'd say that if you postulate a mysterious, separate, additional, inherently mental property of consciousness, above and beyond positions and velocities, then, at that point, you have *already* stuck your neck out. To postulate this stuff of consciousness, and then further postulate that it *doesn't do anything*—for the love of cute kittens, *why?*

There isn't even an obvious career motive. "Hi, I'm a philosopher of consciousness. My subject matter is the most important thing in the universe and I should get lots of

funding? Well, it's nice of you to say so, but actually the phenomenon I study doesn't do anything whatsoever."

Chalmers is one of the most frustrating philosophers I know. He does this really *sharp* analysis... and then turns left at the last minute. He lays out everything that's wrong with the Zombie World scenario, and then, having reduced the whole argument to smithereens, calmly accepts it.

Chalmers does the same thing when he lays out, in calm detail, the problem with saying that our own beliefs in consciousness are justified, when our zombie twins say exactly the same thing for exactly the same reasons and are wrong.

On Chalmers's theory, Chalmers saying that he believes in consciousness cannot be causally justified; the belief is not caused by the fact itself, like looking at an actual real sock being the cause of why you say there's a sock. In the absence of consciousness, Chalmers would write the same papers for the same reasons.

On epiphenomenalism, Chalmers saying that he believes in consciousness cannot be justified as the product of a process that systematically outputs true beliefs, because the zombie twin writes the same papers using the same systematic process and is wrong.

Chalmers admits this. Chalmers, in fact, explains the argument in great detail in his book. Okay, so Chalmers has solidly proven that he is not justified in believing in epiphenomenal consciousness, right? No. Chalmers writes:

> Conscious experience lies at the center of our epistemic universe; we have access to it *directly.* This raises the question: what is it that justifies our beliefs about our experiences, if it is not a causal link to those experiences, and if it is not the mechanisms by which the beliefs are formed? I think the answer to this is clear: it is *having* the experiences that justifies the beliefs. For example, the very fact that I have a red experience now provides justification for my belief that I am having a red experience...
>
> Because my zombie twin lacks experiences, he is in a very different epistemic situation from me, and his judgments lack the corresponding justification. It may be tempting to object that if my belief lies in the physical realm, its justification must lie in the physical realm; but this is a *non sequitur.* From the fact that there is no justification in the physical realm, one might conclude that the *physical* portion of me (my brain, say) is not justified in its belief. But the question is whether *I* am justified in the belief, not whether my *brain* is justified in the belief, and if property dualism is correct than there is more to me than my brain.

So—if I've got this thesis right—there's a core you, above and beyond your brain, that believes it is not a zombie, and directly experiences not being a zombie; and so its beliefs are justified.

But Chalmers just *wrote all that stuff down,* in his very physical book, and so did the zombie-Chalmers.

The zombie Chalmers can't have written the book *because* of the zombie's core self above the brain; there must be some entirely different reason, within the laws of physics.

It follows that even if there *is* a part of Chalmers hidden away that is conscious and believes in consciousness, directly and without mediation, there is also a *separable subspace* of Chalmers—a causally closed cognitive subsystem that acts entirely *within* physics—and this "outer self" is what speaks Chalmers's internal narrative, and writes papers on consciousness.

I do not see any way to evade the charge that, on Chalmers's own theory, this separable outer Chalmers is deranged. This is the part of Chalmers that is the same in this world, or the Zombie World; and in either world it writes philosophy papers on consciousness *for no valid reason.* Chalmers's philosophy papers are not output by that inner core of awareness and belief-in-awareness, they are output by the mere physics of the internal narrative that makes Chalmers's fingers strike the keys of his computer.

And yet this deranged outer Chalmers is writing philosophy papers that *just happen* to be perfectly right, *by a separate and additional miracle.* Not a logically necessary miracle (then the Zombie World would not be logically possible). A physically contingent miracle, that *happens* to be true in what we think is our universe, even though science can never distinguish our universe from the Zombie World.

I think I speak for all reductionists when I say *Huh?*

That's not epicycles. That's, "Planetary motions follow these epicycles—but epicycles don't actually *do* anything—there's something else that makes the planets move the same way the epicycles say they should, which I haven't been able to explain—and by the way, I would say this even if there weren't any epicycles."

According to Chalmers, the causally closed system of Chalmers's internal narrative is (mysteriously) malfunctioning in a way that, not by necessity, but just in *our* universe, miraculously happens to be correct. Furthermore, the internal narrative asserts "the internal narrative is mysteriously malfunctioning, but miraculously happens to be correctly echoing the justified thoughts of the epiphenomenal inner core", and again, in *our* universe, miraculously happens to be correct.

Oh, come on!

Shouldn't there come a point where you just give up on an idea? Where, on some raw intuitive level, you just go: *What on Earth was I thinking?*

Humanity has accumulated some broad experience with what correct theories of the world look like. *This is not what a correct theory looks like.*

"Argument from incredulity," you say. Fine, you want it spelled out? The said Chalmersian theory postulates multiple unexplained complex miracles. This drives down its prior probability, by the conjunction rule of probability and Occam's Razor. It is therefore dominated by at least two theories which postulate fewer miracles, namely:

- Substance dualism:
  - There is a stuff of consciousness which is not yet understood, an extraordinary super-physical stuff that *visibly affects* our world; and this stuff is what makes us talk about consciousness.
- Not-quite-faith-based reductionism:
  - That-which-we-name "consciousness" happens *within* physics, in a way not yet understood, just like what happened the last three thousand times

humanity ran into something mysterious.
    - Your intuition that no material substance can possibly add up to consciousness is incorrect. If you *actually* knew *exactly* why you talk about consciousness, this would give you new insights, of a form you can't now anticipate; and afterward you would realize that your arguments about normal physics having no room for consciousness were flawed.

Compare to:

- Epiphenomenal property dualism:
    - Matter has additional consciousness-properties which are not yet understood. These properties are epiphenomenal with respect to ordinarily observable physics—they make no difference to the motion of particles.
    - *Separately*, there exists a not-yet-understood reason *within normal physics* why philosophers talk about consciousness and invent theories of dual properties.
    - *Miraculously,* when philosophers talk about consciousness, the bridging laws of *our* world are exactly right to make this talk about consciousness correct, even though it arises from a malfunction (drawing of logically unwarranted conclusions) in the causally closed cognitive system that types philosophy papers.

I know I'm speaking from limited experience, here. But based on my limited experience, the Zombie Argument may be a candidate for *the most deranged idea in all of philosophy.*

There are times when, as a rationalist, you have to believe things that seem weird to you. Relativity seems weird, quantum mechanics seems weird, natural selection seems weird.

But these weirdnesses are pinned down by massive evidence. There's a difference between believing something weird because science has confirmed it overwhelmingly—

—versus believing a proposition that seems downright deranged, because of a great big complicated philosophical argument centered around unspecified miracles and giant blank spots not even claimed to be understood—

—in a case where *even if you accept everything that has been told to you so far,* afterward the phenomenon will still seem like a mystery and still have the same quality of wondrous impenetrability that it had at the start.

The correct thing for a rationalist to say at this point, if all of David Chalmers's arguments seem individually plausible, is:

"Okay... I don't know how consciousness works... I admit that... and maybe I'm approaching the whole problem wrong, or asking the wrong questions... but this zombie business *can't possibly be right.* The arguments aren't nailed down enough to make me believe this—especially when accepting it won't make me feel any less confused. On a core gut level, this just *doesn't look* like the way reality could *really really* work."

But this is not what I say, for I don't think the arguments are plausible. "In general, all odd numbers are prime" looked "conceivable" when you had only thought about 3, 5,

and 7. It stopped seeming reasonable when you thought about 9.

Zombies looked conceivable when you looked out at a beautiful sunset and thought about the quiet inner awareness inside you watching that sunset, which seemed like it could vanish without changing the way you walked or smiled; obedient to the plausible-sounding generalization, "the inner listener has no outer effects". That generalization should *stop* seeming possible when you say out loud, "But wait, I am thinking this thought right now inside my auditory cortex, and that thought can make my lips move, translating my awareness of my quiet inner listener into a motion of my lips, meaning that consciousness is part of the minimal closure of causality in this universe." I can't think of anything else to say about the conceivability argument. The zombies are dead.