

COULD DANIEL DENNETT BE A ZOMBIE?

by Mike Kearns

Could Daniel Dennett be a zombie?

The way he tells it, you'd almost have to say yes. For he has been kind to zombies in his recent writings.¹

The precursors of (philosophers') zombies were machines that had to be in another room, communicating by teletype, in order to attempt the feat of passing for humans.² Then full-fledged zombies took the philosophical stage. They were supposed to be *just like us*, but without consciousness.

The "just like us" part is tricky, and is the locus of Dennett's favoritism. Originally it seemed to mean, humanoid creatures whose behavior is indistinguishable from ours. Digging deeper, it meant creatures whose wiring is such that they reach the same results as we do when processing the data of existence, and act like we do as a result. The idea was that a creature could do exactly that, without being conscious. It could be computationally equipped so as to be behaviorally indistinguishable from us, but there just wasn't any "light" on, there was nothing it was like to be that creature, in Nagel's phrase.

The pivotal observation in this zombie thought experiment was that third-person science wouldn't be able to tell a zombie from a real human. Because by definition, in all external ways they behave the same way. The zombie cries when it hears a sad song. The zombie gets the right answer to a difficult math problem – sometimes. The zombie says it believes that it isn't a zombie. With this established, a couple of moves offered themselves. One was to say that we who are conscious are superior to zombies, so third-person science is missing a crucial element of the universe (Dennett calls this the *Zombie Hunch*). The other was to say that zombies are just as good as us, therefore third-person science isn't missing anything important.

Kindness to zombies has gone a little too far though. We find Dennett recently portraying a certain zombie as fervently believing that he is not a zombie, and Dennett acts as if this imputation is licensed by the mere definition of a zombie. (SD 48) On similar grounds we also find him saying that a zombie is indistinguishable from you or me in its neurological wiring (SD 15). Some might claim that that was built in to the concept all along, but I would say it begs a very big question.

Many of us believe that the mind *is* the brain, in some sense. We celebrate the good news that what's going on in the brain fully accounts for what's going on in the mind. There are two camps as to what *kind* of good news this is. The Reducers of Consciousness – call them ROCs – feel that this is good news because it means that the mind is "only" the brain. All the magic of the mind is really just neuron weather. Anything that seems like it couldn't be accounted for in the neuronal universe must be unreal. The Defenders of Consciousness – the DOCs, of whom I am one – reply that no, the good news is that everything we know about the mind now forcibly expands our knowledge of the brain. We don't have to jettison the rich stores of shimmering data – sometimes called phenomenology – we've gathered about our mental life. We now know

¹ Much of this paper was sparked by my reading Dennett's *Sweet Dreams* (2005), hereafter SD.

² These pre-zombie machines were proposed by Alan Turing in his 1950 *Mind* article, "Computing Machinery and Intelligence".

that a complete theory of the brain will *include* all this bounty. So the mind's homestead in the brain has raised the stakes in brain-description, not lowered the stakes in mind-description.

Consciousness, far from being an inconvenient, resistant datum that we're saddled with in our study of the brain, is the thing we can use to unlock the brain's mysteries. Because we are all housed in brains, we have the advantage of knowing in advance, from the inside, what a portion of the brain is up to. We can use this partial key to break the whole code.

When the brain is understood, we DOCs declare, it will become clear that what goes on in the brain does not explain conscious thought, does not reduce conscious thought to something less. No, what goes on in the brain *is* conscious thought as we always knew it. Like a valuable country house finally ending up in the right hands, the riches of consciousness are finally granted to the brain; and that does not reduce our understanding of consciousness – it challenges our understanding of the brain.

Let me give one small example. We have all had the experience of comprehension. It's a pretty routine experience but sometimes, when it emerges from difficulty, it can be downright dramatic. Sometimes we stare at a puzzle, or a bunch of facts, or a statement in a foreign language, or a list of procedures we've just been trained on at a new job – and we just can't make sense of it. We can rehearse its parts but we can't make them add up to anything. Then suddenly we get it. We see the solution to the puzzle. We see the implication of the facts: the wife's masseur must have killed the husband. We see that the statement in Latin means "The advancing army lost heart when they reached the river." We see that the procedures our supervisor has just taught us make sense because the protagonists of this database are the product numbers, not the client sites.

That's my example. The dawning of comprehension. A thrilling thing, an experience. We now know that the dawning of comprehension is something that goes on in the brain. (Where *else* could it happen?) Some aroused state of neurons must *be* the dawning of comprehension. The brain pulls this off.³ So the brain, having had the distinctions of conscious experience conferred upon it by mind-brain identity theorists, is shown to be a wondrous thing – even more wondrous than we already thought it was. Tomorrow or the next day, scientists will be observing a brain and say, "There goes a wave of insight." (Notice that the forgoing examples of comprehension all involve a reference to something outside the physical boundaries of the subject. We will return to this point later, for the deceptively simple feat it implies is a defining, and jaw-dropping, endowment of consciousness.)

Zombies, on the other hand, don't comprehend anything. Because they aren't conscious. So they *couldn't be* indistinguishable from us in their neurological landscape. Zombies may have something that works sort of like a brain, but they can't have brains, because if they did, they would have consciousness. (A brain like ours is sufficient to

³ Some philosophers may want to correct us by saying that brains and states of brains aren't conscious of anything, don't have insights, etc.: it is the *possessors* of the brains that are properly described in these ways. Though such a warning demonstrates a laudable respect for the way people talk, it merely blunts the theoretical questions that concern us – that is, unless such philosophers hold that consciousness resides in the feet. See for example Dretske, "What good is consciousness?" in the Function of Consciousness section of <http://consc.net/online1.html>.

generate consciousness, after all.) So zombies don't have brains. Nor do they have beliefs, feelings, or doubts. At least that is what I believe, because I am a DOC. So what *do* zombies have? Let us stick our necks out very far and hazard the guess that zombies have computers in their heads; reliable intelligence has it that this is what the ROCs have in mind.

We return now to the question, can third-person science tell a zombie from a human? Now that we've recognized peculiarities inside the zombie head, the question looks a little different. If third-person science opens up the zombie's head and looks inside, they'll see that it isn't a human, and the jig will be up. What about if they *don't* look inside? Well then they won't know. Except...except... something still doesn't sit right. We still suspect that zombies will be found out, that they won't pass muster. We start to wonder whether, without comprehending anything, zombies could really behave just like humans.

Liking to ask questions that confound common sense, philosophers have asked: does consciousness confer any advantages that could explain why evolution would so carefully have nurtured it? And if not, do we really have to trouble ourselves with consciousness? Common sense responds, "Try driving behind a guy who's asleep, and see if consciousness has any advantages."

Philosophers (ROCs in particular) reply: "Oh, we didn't mean *that* kind of advantage; or if we did, we didn't mean conscious as opposed to being asleep. A zombie keeps his eyes on the road, and his visual computational center contributes to his driving safely, which his prudential computational center mandates." And the ROCs continue: "What we meant was, does a perfect driver who is experiencing the trip have any survival advantage over a perfect driver who is not experiencing anything? And our answer is No."

Well, right – *if* that was the choice put before Evolution. Makes you wonder, why are there not more computers in our planet's illustrious history? Why all these conscious beings? According to a neglected screenplay called *Valley of the Lost Microchips*, long ago a fertile vale in South America was in fact riddled with computers, dumped there by an alien civilization that had grown weary of technology. The problem was, they just sat around. They were of no use to anyone until the dinosaurs came along and ate them.

Suppose you were Evolution and you were given a choice:

A. you can develop machines that compute with amazing rapidity and accuracy, but they aren't aware of anything. Nothing means anything to them. They don't believe anything. They don't comprehend anything.

or

B. you can develop creatures that are conscious, that are aware of objects in the world around them. They believe things. Later, they comprehend things: many things. Before too long, they communicate with each other using symbols whose meaning they understand.

Which would you choose? If I were Evolution I'd say, "Well, Option B sounds a lot quicker. With the help of awareness (and other concomitant features) this line of development can move along pretty fast. And the resulting creatures sound like they have a lot more survival prowess. Intra-species linguistic communication especially appeals to me." And Evolution might add, "Option A sounds okay too, I do believe that computers

could blanket the world if we could get them started, but that won't happen without help. Wait! I've just thought of a way to develop them. I'll choose Option B and then those conscious creatures will invent computers, and eventually those computers may even be developed into zombies, though even I may not be able to pull off this last step."

Is a zombie really possible? Dennett himself has expressed a desire to stop talking about zombies, and has argued in some contexts that they may be less viable, in terms of logical or physical possibility, than one might have thought by reading his works.⁴ It seems that he only entertains them for the sake of argument, and then never at home. (His favorite use of them is to meet them at his local pub, slip them a consciousness mickey, and then claim that consciousness is not worth much, since zombies' company is just as good as ours.)

A juncture in which Dennett is less favorable to zombies is the one where, seeking to lull the Defenders of Consciousness into a tranquil unmindfulness, he portrays himself as a DOC-manqué. In this mood he affects a warm regard for consciousness, assuring us that all of its fine features will be encompassed in his third-person account. By practicing an admirable "heterophenomenology", by embracing the "Intentional Stance" with all the passion that behaviorism allows, he will produce a theory that confines itself to what third-person scientists and their unexpected allies the Martians are able to know – yet that theory will not neglect any nuance of consciousness worth mentioning.⁵ (How far this claim falls short we will see in a moment.)

When he takes this stance, Dennett's belated abandonment of the "unfortunate" topic of zombies is not hard to explain (SD 150), for the sort of consciousness his theory allows is clearly provable of zombies, and that would contradict his definition of them. If he persisted in saying that there could be beings indistinguishable from us by third-person science yet not conscious, he would effectively unmask the counterfeitness of his claim to be honoring consciousness in his third-person theory.

In spite of his claim to have "long claimed" that the conceivability of zombies is only apparent (SD 15), one suspects that it is *non-zombic* humans that Dennett finds if not inconceivable, then unendurable. Consider the convenience of zombies to Dennett's enterprise. Zombies talk and act like they have consciousness, and they perform perfectly as subjects of heterophenomenology, thus meeting all the demands Dennett thinks an adequate theory of consciousness would make of them. Again to zombies' credit, they are not conscious "in any of the tendentious ways much discussed of late by philosophers" (SD 27). Zombies are *perfect* for Dennett! If the human world turned out to be populated only by them, his troubles would be over. Non-zombic humans, by contrast, are a nightmare for Dennett.

When Dennett pretends to renounce zombies, he means that consciousness is nothing beyond that which can be tested behaviorally and objectively, so since zombies

⁴ See SD, pages 13, 80, 92, and 150.

⁵ Dennett introduces his Martian scientists at the start of his ambitious chapter, "A Third-Person Approach to Consciousness" (SD 25). These investigative allies seem to be recruited to lend an extra degree of distance from human subjectivity, but it becomes clear that Dennett wants to have it both ways with them, for they are just tricked-up zombies. He starts by crediting them with sense organs, beliefs and "the knack of adopting the *intentional stance*", then claims not to be presupposing that they are conscious, then goes for the haymaker, saying they are zombies "without a trace of...real consciousness" (SD 26-27). Later in the book he denies that such a formulation is even coherent (SD 92, 150).

pass these tests, they can't be said to be unconscious – yet that is part of their definition. So they are a contradiction. In other words, zombies are inconceivable because a being that is "behaviorally, objectively indistinguishable from a conscious person" just doesn't deserve in Dennett's eyes to be called unconscious. (SD 150)

It is particularly ironic to find Dennett quoting Nagel with approval as agreeing with him about zombies (SD 15). When Nagel is troubled by the apparent conceivability of zombies, he means, of course, that if brain states like ours could function just as well in a zombie, that would seem to entail that they can't be the seat of consciousness, which leaves that valued thing homeless. Nagel wants zombies to be inconceivable because he wants to leave open the possibility that consciousness as he describes it might reside in the brain; Dennett claims (less ingenuously) to dismiss zombies for the opposite reason: he wants to call them conscious.

Lest anyone doubt that Dennett's real goal remains that of excluding consciousness from the ultimate inventory of the universe, I offer the following pieces of evidence:

First, consider these premises, all espoused by Dennett in *Sweet Dreams*:

- a) zombies do not have consciousness
- b) zombies have brains just like ours
- c) if consciousness exists in our universe, brains like ours are sufficient to generate it.

The inference is clear.

But it is hazardous to argue a person's belief from other beliefs he is known to hold; so some *direct* evidence seems in order, showing that Dennett, not unlike traditional behaviorists, believes the universe can be fully described without mentioning anything conscious.

The following passages should do the trick:

"All the work...[of consciousness]...must be distributed among various lesser agencies in the brain, none of which is conscious... the Subject vanishes, replaced by mindless bits of machinery unconsciously executing their tasks." (SD 69-70)

"...those who work on answers...are not leaving consciousness *out*, they are explaining consciousness by leaving it *behind*." (SD 144)

"Heterophenomenology...is a reasoned, objective extrapolation from...the behavior of subjects, including especially their text-producing or communicative behavior..." (SD 149)

It still bears asking, why have a long line of otherwise hale and resourceful thinkers like Daniel Dennett wanted so badly to exclude consciousness from their ultimate inventory of the universe? There *is* a reason, an honorable, powerful one. The gist of it is this: the ROCs are afraid that consciousness can do something that brain science won't be able to handle. Something just too peculiar, something no amount of third-person investigation will even give *access* to. If they leave consciousness

unscathed, science will end up humiliated by explanatory inadequacy. So in defense of the brain and of science, they seek to deny consciousness.

We refuse to do that. We relish consciousness in all its stubborn oddness, and in the rest of this paper we will assume that "leaving it behind" is not an option. That still leaves us with the question: will we ever be able to explain consciousness in its undiluted form? What gives brains their unique gift? What is it about them that engenders full-blown consciousness as we know it?⁶

Well, that's what science needs to find out, isn't it? And here we encounter another fork on the DOC road. The pessimistic DOCs (McGinn & the Mysterians) say science never will find out. Like the ROCs, they say that the ore of Experience could never be found in Neuron Mountain (they differ from them in refusing to jettison that ore). But we, the optimistic DOCs, think otherwise. We think that some day science *will* pinpoint the peculiarities of brains that give rise to consciousness.

It may be that something about the wiring diagram of the brain will turn out to guarantee consciousness. If such a point is grasped it will surely be among science's greatest triumphs. The formula which describes the exact interactions that allow billions of neurons to create what we call experiences, will no doubt be brain-boggling. It may be that once it is understood, we will in principle be able to create an array of non-organic switches which conjures consciousness. Or it may be that the organic stuff the brain is made of is also essential to its success (possibly only in the sense that there is no other practical way to create the desired wiring diagram). Other conditions may also be unearthed, without which the set is not sufficient.

In any case, at this point scientists will have the formula for consciousness. (Or rather, a formula. It may not be the only possible one in the cosmos.) In a common scientific sense they'll know "why" a given creature is conscious: it's because the stated set of conditions is met. So they will have a comfort level around consciousness that they never had before.

But some (like McGinn) would say that in another important sense of 'why', they still won't know why that set of conditions brings about consciousness. My response is that if brain states really are identical to conscious states, *there is no why*. Why is this brain state me loving the melody of "Yesterday"? Well, it just is. There is nothing to explain.

Going down the ladder one rung, will there be a chance of explaining brain states in terms of something else? Using William Seager's delightfully simple rules of naturalization⁷, we can say that the question is not

"Can consciousness be explained in terms of brain states which do not essentially involve consciousness?"

⁶ For this discussion I confine myself to the *human* brain. Interesting issues arise as to whether science, once it gets a handle on consciousness, will be able to detect consciousness in other living creatures such as bats and cats. I expect the answer is yes.

⁷ See William Seager, "Real Patterns and Surface Metaphysics", in *Dennett's philosophy: A Comprehensive Assessment*, edited Ross, Brook, and Thompson, MIT, 2000, page 96. Seager affirms that X has been naturalized iff X "has been explained in terms of Something Else", the "Something Else does not essentially involve X", and the Something Else is "*properly* natural".

but rather

"Can brain states, understood as conscious, be explained in terms that don't involve brain states?"

And the answer, this DOC believes, is no. To see why, consider that the brain is fairly seething with human thoughts, feelings, and more generally, experiences. Human experiences carry aboutness in a way unlike that achieved by words on a page or internal states of today's computers: namely, *intrinsically*. Therefore, brain states do too.⁸

Suppose I am watching a *West Wing* rerun in which President Bartlet looks at a set of fine pens in a desk drawer and thinks about his beloved secretary, Mrs. Landingham, who recently died in a car crash and who used to slip one of those pens into his pocket every morning. The sadness of the scene makes me tear up. Any attempt to describe (or account for) the related brain states without mentioning President Bartlet, Mrs. Landingham, and my perception of the scene as *sad*, will simply be misreporting them: will leave out something essential to the intrinsic content of the brain states. When I say "intrinsic" I mean to imply that a brain could not be in exactly this sequence of states without the cosmos containing an experience of Bartlet's plight as sad. Any attempt to describe brain states without mentioning their units of content will miss the boat. (More about the aboutness of brain states in a moment.)

Our scientific understanding of brain states, then, needs to rise to the complexity, richness, and "imbued with meaning"-ness of experiences. One aspect of this gain in insight is that scientists' structural individuation of ongoing brain processes needs to be guided by the individuation of experiences. As Kohler eloquently pointed out, one of the most striking features of human experiences (and therefore of brain processes) is their seamless integration of material from disparate "departments" of perception, emotion, and concept into gestalts.⁹ Instead of indulging Dennett's drive to break brain states up into ever "lesser" parts, science will do well to chart the brain's talent for integration of materials into ever-more-complex wholes.

In connection with naturalization a la Seager, I said that brain states, understood as conscious, will not be explained in terms that don't involve brain states. This does not deter me from saying that I am an optimistic DOC. My reason: I believe that brain states will be naturalized in a way omitted from Seager's presentation. It seems to me that there is another way X can be naturalized, besides Seager's way of being explained in terms of "Something Else" that is "properly natural". That other way is: X can be shown to be properly natural itself. And I believe that brain states, understood as conscious, will inevitably be inducted into the Properly Natural hall of fame.

At that point there will be nothing left to explain in terms of something else, just a wondrous phenomenon left to appreciate.

Whatever the formula for consciousness turns out to be, we DOC optimists think its day of discovery will come. On that day, we believe that scientists will at last know what separates brains from computers. (Of course, it may turn out that we can construct

⁸ I use brain states as convenient shorthand for some more laborious phrase such as "ongoing processes of portions of the brain".

⁹ See, for example, Kohler, *Gestalt Psychology*.

thinking machines that incorporate those awe-inspiring brain conditions. If we call them Komputers, it will be true that Komputers have consciousness. But that won't prove the ROCs were right about zombies rivaling us in their prowess, because zombies won't be able to have Komputers in them, since zombies aren't conscious.)

Will the ROCs get a Pyrrhic victory? On that fine day when consciousness is "solved" (albeit at the price of admitting that brain states cannot be explained in terms of something else), will "third-person" scientists at least be able to announce that there is now nothing left about consciousness that eludes their grasp? The sad answer is No.

The reason for this has been linked to the term 'qualia'. What the color blue looks like to me is said to be a *quale*, something which no one but I can know, something I am directly acquainted with as they used to say.¹⁰ The negative intuition about third-person science, then, goes something like this: no matter how thoroughly a scientist might poke around in my brain while I am experiencing blueness, she will never get to anything that shows her the quale blue. My brain is experiencing it, but she can't find it *in* my brain; and if, like Mary the color scientist in the famous thought experiment,¹¹ she does not experience colors herself, she will not have *any* way of accessing that quale. If, however, she *does* experience colors and if brain science has advanced enough, she may be able to mount an argument – based on consonances between her brain states and mine – that she at least knows the same quale blue that I do. For it would be odd if Nature allowed two different color experiences to spring from the same neural formula.

Yes, I am supposing that the neural formula for blue is different from the neural formula for red. And that there *is* a neural formula for blue. If the mind is the brain, something like this must be true. There must be a state of a portion of the brain such that nothing in the universe can be in that state without an experience of blue taking place. And that must have been true all along, true long before there was anybody around to see anything. Taken seriously, this claim taunts the scientific imagination. It is hard to imagine that a set of neurons, simply by being in a certain state – however rich in numbers, organic, convoluted, and interactive – can force the universe to contain an experience of the quale blue, while a set of neurons¹² in a different state can force the universe to contain an experience of the quale red. We wonder, what in the world is that team of neurons up to; what are they *doing* that brings about such an astonishing *frisson*? But we turn to science in good faith, honoring it with the belief that it will be equal to answering this question, at least in the sense of telling us exactly *what* those neurons have to do to generate the experience. Third-person scientists, however, still won't know what experience they're explaining, unless they have it themselves.¹³

¹⁰ Moore, Russell and others in fact *defined* sense-data (and their properties, like blueness) in terms of our direct acquaintance with them, with the result that it became true *a priori* that sense-data could not be ordinary objects, and instead formed a sort of wall between us and the world.

¹¹ Dennett presents Johnson's 1982 thought experiment in Chapter 5 of SD, memorably calling it an intuition pump.

¹² Could this latter set of neurons be the same set? In principle it could turn out that there are specialized neuron sets whose job it is to produce the experience of blueness, and others that produce redness; or it could turn out that the same set of neurons can produce either experience by being in the right respective state.

¹³ That is why Dennett's notion of zombie Martians carrying on heterophenomenology is so stupefying. Only conscious beings can study *anything* – including consciousness. For Dennett's team to argue that consciousness doesn't exist is like a team of investigators claiming to prove that arsonists don't exist, while lighting illegal fires to better view their subjects.

What makes qualia so popular with philosophers is that, as good luck would have it, some conscious beings are denied access to the qualia enjoyed by some other conscious beings. This inaccessibility provides an extreme example to aid the argument that a description of the brain, no matter how complete, will still leave something out. Conscious beings have a tendency to smuggle things that only conscious beings can know into their study of other conscious beings. So it's helpful when we can adduce a case where something a particular scientist cannot know (say, the experience of blueness) is known by the person whose brain she is examining.

I think, however, that this inaccessibility is a red herring. I think that what is most noteworthy about qualia can be best appreciated by a scientist who *does* possess the experience in question. Two paragraphs ago, my knowledge of what the experience of blue is like was exactly what fueled my sense of how odd it is that a certain brain state could possibly generate it. My wonder was of the form: "one would never guess, examining that brain state, that it was generating *this* experience." I can poke around in your brain but I will never find "blueness", *even if I myself do have the experience of blue*. So lack of access isn't really the point.

In any case qualia are not as astounding, when considered as products of the brain, as the thing which they facilitate.¹⁴ It is just hard to appreciate this, because *all* conscious beings have access to the thing that is more astounding than qualia. This thing may be the most outrageous trick that the universe has ever pulled off.

In order to pinpoint it we want to say an incoherent thing: we want to say, a scientist who wasn't conscious could never find out by looking at my brain that I am. (Just as a scientist who didn't have blue could never discover by looking at my brain that I am experiencing blue.) But the problem is, *all* scientists are conscious. And when Dennett says "third-person science" he is somehow trying to trade on the idea of a scientist who isn't conscious, who can only observe others from without; but that too is malformed, because *all* humans can only observe others from without. So what work is the term "third-person" doing?

And here we have to resort to metaphor; but we should not be leery of that, because metaphor is one of the time-honored tools in the difficult enterprise of describing the ineffable. We have been strolling along its shiny strand, but now we must step into the waves.

What is it that "third-person" science can't find in my brain, even though it is my brain that is contributing it to the universe?

In a word, it is my world. Or my having of a world.
For the brain comes with a world. The world as experienced by me.

The brain doesn't go: Judging by these signals, there must be a world out there.
The brain goes: Look, a world.

¹⁴ Metaphysically speaking, qualia are to objects in the external world what bandages were to the invisible man: they are what allow that world to be known by us.

That was the great, mind-boggling breakthrough. When a brain went: Get a load of that *thing*.

That was the birth of intentionality.

Language can mean something because the brain has a world.

Brain states don't need meanings assigned to them, the way language does. Brain states don't need interpretations. When the brain is in a certain state, I get a world.

The universe had us at hello.

The world the brain comes with is not built in, but built out. Not a world inside the brain, but a world outside the brain. Not just any world. The real world. The real world as experienced by the possessor of that brain.

And that is why Daniel Dennett cannot be a zombie. Zombies don't have a world. But Dennett does.