The Zombie Threat to a Science of Mind

(Published in May/June edition of 'Philosophy Now')

For the last five hundred years or so, physics has been doing extraordinarily well. More and more of our world has been captured in its explanatory net, from the formation of planets and stars, to the nature of space and time, to the very basic constituents of matter that make us up. There's a long way to go: our best theory of the very big, i.e. general relativity, is inconsistent with our best theory of the very small, i.e. quantum mechanics. But many look forward to the day when physicists will resolve these niggling issues, and will present the public with the holy grail of science: the Grand Unified Theory of Everything. The hope of many philosophical inclined scientists and scientifically enthused philosophers is that this theory will explain the existence and nature of everything there is. Let us call this kind of view 'physicalism.'

Physicalism is a grand and ambitious project, but there is a thorn in its side: consciousness. The qualities each of us encounters in our inner conscious experience – the feeling of pain, the sensation of biting into a lemon, what it's like to see red – stubbornly refuse to be incorporated into the physicalist's all-encompassing vision of the universe. Consciousness seems to be the one bit of left-over magic that refuses to be naturalised. And it's all the fault of the zombies.

I'm not talking about Hollywood zombies, the lumbering, semi-decayed, undead creatures that the rest of this edition is taken up with. In philosophy of mind 'zombie' is a technical term for a rather specific kind of creature that features large in philosophical thought experiments.

A *philosophical zombie*, as opposed to a Hollywood zombie, is a *physical duplicate of a human being that lacks consciousness*. A zombie version of you would walk and talk and in general act just like you. If you stick a knife into it, it'll scream and try to get away. If you give it a cup of tea it'll sip it with a smile. It uses its five senses to negotiate the world around it in just the way you do. And the reason it behaves just like you is that the physical workings of its brain are indiscernible from the physical workings of your own brain. If a brain scientist cut open the heads of you and your zombie twin, and poked around inside, they would be unable to tell the two apart.

However, your zombie twin has no inner experience; there is nothing that it's like to be your zombie twin. Its screaming and running away isn't accompanied by a feeling of pain. Its smiles are not accompanied by a feeling of pleasure. Its negotiation of its environment does not involve

a visual/auditory experience of that environment. Your zombie twin is just a complex automaton mechanically set up to behave just like you. The lights are on but nobody's home.

Nobody believes zombies really exist. But some philosophers think they're *possible*. Compare to flying pigs. There aren't any flying pigs. But there is no contradiction in the idea of a flying pig, in the way there is a contradiction in the idea of, say, a square circle. If things had been very different – if pigs had evolved with wings, or if the law of gravity had been much weaker – there could have been flying pigs. In contrast no matter how weird and whacky the universe had turned out, there could not have been square circles; the very idea is incoherent. So although neither flying pigs nor square circles exist, flying pigs have the advantage over square circles of being possible.

So although no philosophers think that zombies are real, some think that zombies, like flying pigs, are possible. If the universe had turned out very differently, perhaps if the laws of nature had been radically different, there could have been zombies.

Sometimes possibilities matter

Now you might think: why should the physicalist care about possibilities? Physicalism is a view about the *real world*. Why should the mere possibility of some weird non-existent creature have any bearing on a serious, scientifically supported theory of the actual world? The trouble is that sometimes what's possible has implications for what's real. And there is a broad consensus amongst philosophers that *the mere possibility* of zombies if inconsistent with physicalism. I will now try to explain the reasoning behind this consensus.

It seems that if we want the physical sciences to account for consciousness, then we're going to have to identify states of conscious experience – the feeling of pain, the experience of biting into a lemon – with *brain states*. Neuroscience is far from complete, but at our current level of knowledge there seems to be no reason to think that the kind of things neuroscience deals with, things like neurons and neurotransmitters, could not in principle be completely explained in

terms of the properties of their physical constituents. So if we can identify conscious states with brain states, and we can explain brain states in terms of their physical constituents, then we will have thereby explained consciousness itself in terms of the physical bits of brains.

The trouble is that it follows from the logic of identity that: if zombies are possible, conscious states cannot be identified with brain states. For X and Y to be identical is for X and Y to be *one and the same thing*. Cliff Richard is identical with Harry Webb. That is to say, we do not have two people – Cliff *and* Harry – we just have one person with two labels. Similarly, to say that the feeling of pain is identical with c-fibres firing is to say that *pain* – the thing you feel when a knife is stuck in you – and *c-fibres firing* – the thing the brain scientist sees when she looks in your head after you've had a knife stuff in you – are one and the same thing. We don't have two things – pain *and* c-fibres firing – but one thing with two labels.

Now, if X is identical with (is one and the same things as) Y, then X can't possibly exist without Y. Think about it. Cliff Richard *just is* Harry Webb, so of course the two can't exist apart. After all, there aren't two people to separate; Cliff *is* Harry. Not even God could pull Cliff Richard away from Harry Webb. Similarly, if pain is identical with (is one and the same thing as) c-fibres firing, then not even God could pull them apart, as there aren't two things to separate: pain *just is* c-fibres firing. But in your zombie twin, c-fibres firing *does* exist without pain; your zombie twin has all your brain states in the absence of any of your conscious states. Therefore, if zombies are possible, then your brain states could exist without your conscious states, and therefore your brain states aren't identical with your conscious states. If zombies are possible, physicalism must be false.
We can put the 'zombie argument' against physicalism as follows:

1. Zombies are possible.
2. Therefore, human brain states could possibly exist without human conscious states.
3. Therefore, human brain states cannot be identical with human conscious states.
4. Therefore, physicalism is false.

Philosophical resistance to zombies

As I said above, the fact that the possibility of zombies is inconsistent with physicalism is the most uncontroversial part of the zombie argument against physicalism. But there are a couple of ways in which physicalists try to resist the conclusion that zombies are possible. Old school physicalists argue that the very idea of a zombie is incoherent. Upon first reflection, it might seem that the notion of a zombie makes sense. But when you really think about it, claim the old school physicalists, zombies are no more coherent than square circles; the very notion of a zombie involves some sort of subtle contradiction.

The trouble with this old school strategy for getting rid of zombies is that it inevitably involves some kind of *behaviourist* analysis of our mental concepts: to suppose that someone is in pain is

just to suppose that they are responding to bodily damage with avoidance behaviour. If to suppose that someone is in pain is just to suppose that they are behaving in a certain way, then of course the idea of a zombie that behaves just like someone in pain but doesn't really feel pain is going to be incoherent. But most philosophers of mind these days find such behaviouristic account of our mental concepts quite implausible. Of course I *know about* my elderly aunt's pain because of how she's behaving, but the behaviour itself is not what concerns me. Rather I'm concerned about my aunt's *inner feeling of pain*, which her outer behaviour reveals to me.

Since the 1990s, a more popular anti-zombie strategy has been to concede that there is no contradiction in the idea of a zombie, but to try to block the move from zombies being coherent to zombies being genuinely possible. This strategy is made more plausible if a case can be made that there are other scenarios which can be clearly and coherently imagined, and yet turn out to be impossible. Some suggest the scenario of water having some chemical composition other than H_2O as such a case. This seems to be a perfectly coherent scenario: we can imagine scientists discovering that water is composed of something other than H_2O molecules. And yet, if water is identical with H_2O, then, for the reasons discussed above, it is impossible for water to exist without H_2O. And hence the scenario in which water is not H_2O, although it can be coherently imagined, is nonetheless impossible. Most physicalists think the same about zombie scenarios. Although we can coherently imagine human brain states states existing in the absence of human conscious states, such a scenario is impossible, as human brain states just are human conscious states, and so the former could not possibly exist without the latter.

Of course anti-physicalist defenders of the zombie argument have ways of trying to deflate this strategy. Some say that we can't really conceive of water that's not H_2O, but only stuff that superficially resembles water in appearance that's not H_2O. Some point out that if we knew enough about the world (specifically, if we knew that the colourless, odourless stuff that falls from the sky and is found in oceans is H_2O), then we would cease to be able to conceive of non-H_2O water, and yet it seems that no matter how much we learn about the world, we would still seem to be able to conceive of zombies. Dealing with the issues at this point, and trying to get at the exact connection between conceivability and possibility, can get quite technical.
The jury is out on whether or not philosophical zombies are possible. But if they are, then the dream that the physical sciences will one day give us a complete picture of the world can never be realised. This is enough to make many contemporary philosophers more than a little afraid of zombies.