Correlation, Causation, Identity, Subvenience, Emergence, Magic . . .

by **Bob Stone**

David Chalmers sometimes speaks of the need to take consciousness seriously, implying that some philosophers of mind do not. I agree with him, in that at least, and I also agree with him about the hard problem of consciousness: it really is problematic and it's damned hard. So this talk is really about taking Chalmers seriously – the questions he raises, if not necessarily the answers he gives.

1. It is a truth *almost* universally acknowledged that every mental activity, event or state – which includes as a subset every conscious activity, event or state – is accompanied by an equivalent physical event in the body, mostly the brain. Every year neuroscience discovers new **correlations** between mind and brain. Of course, the science is in its infancy, and no one claims to be able to map out in terms of neural activity precisely what went on in Schubert's mind when he was composing – to use an example cited here a few years ago – or what goes on in our minds when we are listening to his music. But the brain has, at the last count, 86 billion neurons, with trillions of connections between them; so, much as some find distasteful the idea of such complex and subtle mental achievements being correlated with, even reduced to, mere physical happenings in the brain, the complexity and subtlety of creativity and feelings is not an issue in principle. The evidence of what we know so far points irresistibly to the assumption that all mental, and therefore conscious, activity is correlated with physical activity. I shall make that assumption in what follows.

Let me take one very specific example of conscious activity to work on. Occasionally, after a stroke or an accident, a patient is left in what is called 'locked-in syndrome', where she is awake (rather than asleep) but completely unresponsive to what is going on around her. Doctors and relatives have no idea whether she is aware of her surroundings or in an unconscious vegetative state. That used to be an insoluble problem, until the arrival of fMRI scanning at the end of the last century. It became possible to read, at least roughly, the activity going on in a patient's brain. In the case of some locked-in patients, it emerged that quite normal neural activity was going on; but the question was whether any of it was conscious. Professor Adrian Owen, a neuroscientist who was studying such patients, had the idea of asking them to imagine doing something, either playing tennis or walking round their house. It had been established that in healthy people, when imagining activities of these kinds (waving arms around or negotiating space), certain brain regions show up strongly in the scanner – the same regions, in fact, as those engaged when they actually do them. So he asked each patient to imagine for 30 seconds that she was playing tennis; remarkably, in a few cases, the patient responded with exactly the same brain activity in the premotor cortex as healthy subjects. The length of time it went on (the full 30 seconds, till he told them to relax) showed that it was not merely an instinctive, possibly unconscious reaction to hearing words the brain recognised, but a deliberate, conscious attempt to follow the doctor's instructions. Later the patients were able to answer yes or no to questions, by imagining either playing tennis or walking round a house, which is accompanied by a quite different brain activity. What is the connection between the conscious experience and the observed brain activity?

2. The first idea is **causation**. This is complicated. Imagine I'm the one in the scanner. It seems intuitively obvious that, within the conscious realm, my understanding the doctor's words, and wanting to oblige him, cause me to imagine playing tennis. In the physical realm, it is fairly clear how sound waves from the doctor's mouth cause events to happen in my ears that trigger neural events that lead to the neural activity shown on the scanner screen – two

parallel causal sequences, each expressed in its own language. But *between* the sequences, we might say that my imagining playing tennis *causes* the neural activity that lights up on the screen; after all, the latter is taken as *evidence* of conscious activity – the whole point of the scanning. And, moving in the opposite direction, a catastrophic accident, or the onset of dementia, may cause me to have conscious experiences of quite a different kind from those I have now – or none at all. It seems difficult to claim that it can be simultaneously true that conscious activity causes the accompanying neural activity, and that neural activity causes the accompanying conscious activity. The two are simultaneous – which is why we say they are correlated – and, while we might claim that one activity *explains* the other in some sense, it is hard to see how one of the simultaneous activities can *cause* the other: we normally assume that, if A causes B, B comes after A.

One entertaining theory is the idea that both sequences, the conscious and the physical, are parallel effects of the same cause – namely God. The Leibniz version of this is, roughly, that, at the beginning of time, God set two clocks in motion: one is the sequence of physical events in the universe, the other is the quite separate sequence of conscious experiences. He ensures that these accompany each other in such a way that that they *appear* to be related.

3. Causation, then, is a difficult idea to sort out. So, on to the second possibility, **identity**. The idea here is that there is only one causal sequence going on, which can be observed from the outside as physical and (partly at least) experienced from the inside as conscious; so my imagining playing tennis *is* the relevant activity of my neurons, albeit seen from a different angle and expressed in different language.

The first problem is the spectre of epiphenomenalism, which goes like this. If the physical sequence – sound waves from doctor, neural activity in my brain – can be explained as a causal sequence by normal physical laws of nature (a system which is considered 'closed' – i.e. not open to any outside non-physical interference), then the accompanying conscious sequence – my understanding the doctor, deciding to imagine playing tennis, and imagining it – cannot be causal in addition; any thought that my deciding causes my imagining is illusory. It is analogous to watching a film, where we pretend to believe that an event that happens on the screen (Rick nodding his head) causes the next one (the orchestra striking up the Marseillaise), though we know perfectly well that what we are seeing is caused by some piece of cinematographic technology. In ordinary life, we are like the ignorant spectators in Plato's cave, treating the moving pictures on the wall as if they are part of a real story, unaware that they are shadows of puppets being operated by puppeteers sitting behind us.

In fact that doesn't seem to me a decisive argument against identity theories; every week neuroscientists discover that something we do, which we had assumed is caused by some conscious decision of ours, in fact started to happen before we became conscious that we had decided it. Less and less of our activity, even our mental activity, is being ascribed to conscious will. But there are more serious objections to the theory that the physical and conscious events are the same events. An early identity theory, proposed by U.T. Place, claimed that, although the conscious experience, say pain, and the neural activity, say the firing of C-fibres, are quite different in conception, we have discovered empirically that they are in fact the same thing. That is contingent identity. Saul Kripke objected that, if two rigidly designated things are identical – that is, two things you refer to directly – their identity must be necessary. It must be part of the essential nature of imagining playing tennis that it is also a certain activity of neurons. He gives as a parallel the discovered identity of water and H_2O, which is a necessary identity. If ever we find on some other planet some water-like stuff that we would intuitively call water, but discover it to be XYZ rather than H_2O, we will know it is not water; if we called it water, we would be misdescribing it. But if we found

beings on another planet who imagined playing tennis with quite different physical things going on in their brains, we would not deny that they were imagining playing tennis; we'd merely think that they had a different physical structure from humans. Therefore, imagining playing tennis is *not* the same thing as the neural activity that accompanies it in human brains.

A variation on Place's so-called type-identity theory is Davidson's token-identity theory (which in fact he applies only to propositional attitudes, but let's extend it to conscious states in general): here it is not that a particular *type* of conscious activity, such as pain, is identical to one type of neural activity, but that any *particular* conscious activity – e.g. my imagining playing tennis now – is the same thing as the relevant neural activity going on in my brain right now. That leaves open the possibility that, when Gareth next door imagines playing tennis, something quite different is going on in *his* brain. It is possible that this gets round Kripke's objection, though I don't myself see how, but it poses another problem. What grounds do we have for claiming that the neural activity going on in some region of my brain is identical with what I'm consciously thinking, unless we have observed that, in my brain and others', the two are correlated. Correlation is by definition between types. If I believe someone is imagining playing tennis, and I observe that she is wearing green shoes, I do not conclude that those two things are identical; it is because the imagining is regularly, typically accompanied by a particular type of neural activity that I judge that, on this occasion, the imagining and the neural activity are linked, and – if I'm an identity theorist – identical.

But the really big problem with identity is this. In the case of water, once we have discovered it is H_20, we can see how. Knowing what we now know about physics and chemistry, we could have predicted that, if you put atoms of hydrogen and oxygen together in a certain way, they would make something wet and transparent. Now is there any way that we could, if we knew everything there was to know about neurons and their behaviour, predict that certain states would give rise to, or be accompanied by, or be identical with certain conscious states, or indeed with any conscious states at all?

4. At this point I will move on to my third possibility, **supervenience**. It is commonly said that higher, or more general, levels of reality are supervenient on more basic levels. For example chemistry is supervenient on physics. That means that, given the physical facts and the laws of nature, the facts about chemistry, though expressed in a different language from those in which we talk about physics, follow inexorably. If the physical facts were different, the chemical facts would be different. Similarly, the facts of biology are supervenient on those of chemistry. And, say some, conscious experiences are supervenient on facts about observable brain activity. There is a crucial distinction which Chalmers makes between two levels of supervenience. *Logical* supervenience is such that, given the physical facts, say, the facts of chemistry *could not have been different* – and that applies to any world with the same facts and laws of physics. But what he calls *natural* supervenience is where we discover that the facts at one higher level do, in this world, happen to supervene on the facts at a lower level, but in another world they might not have done. You couldn't have predicted in advance of the evidence, even from a total knowledge of the facts at the more basic level, that the facts at the higher level would be as they are. In the case of almost all supervenience, he says, the higher-level facts could, in principle at least, be predicted given total knowledge of the more basic; that is logical supervenience. But in the case of conscious experience, which seems to supervene on brain activity, the supervenience – if it is there – is empirically discovered, but is merely *natural* supervenience, because it is *not* then seen to follow inexorably. If a super-intelligent non-human visitor knew absolutely everything there was to know about the workings of the brain, he would still have no idea that conscious experience would accompany some neural states, and not others, or what states they were. How does

Chalmers know that, you may be thinking, given that we really know very little at this stage about the human brain? Isn't it just a matter of waiting until we know more? Well, what we keep learning more of is the observable details: which types of neural activity are accompanied by which types of conscious experience. And in the course of time, who knows, we may have a 100% knowledge of all the correlations – the eliminativist's dream. But it would be no more than empirical knowledge of brute facts; we could never look back and say, "Ah, if we'd had that knowledge of the brain, we could have predicted that conscious states would exist and which conscious states would be accompanied by which brain states."

Why not? Because there is a fundamental divide between subjective conscious experience and observable brain states. All the other levels of reality are observable, albeit with difficulty in some cases; when we wonder how life comes to exist – which some people claim is a parallel – we are observing various things going on in the world and working out how they lead to other observable things that are going on. An alien could do that just as well. But, in the case of conscious experiences, there is – uniquely – something it is like to have them, beyond the omniscient alien's ken.

5. Now to the fourth and final candidate, **emergence**. One neat example of emergence comes from Ann Long on the Society's discussion forum. If you put the right combination of eggs, flour, butter and sugar together and heat them, there emerges something called a sponge. Although, in a sense, there is nothing new that didn't exist before, the existing materials, thus rearranged, now have a new *property* they didn't have before, sponginess. This could not have emerged without the ingredients, but its emergence may well have surprised the first experimenting cook who discovered it. Yet a clever chemist could have worked out in advance, by analysing the ingredients, what they would inevitably be like when heated. The emergence of observed sponginess – a new property with a new word to describe it – from observed unspongy ingredients is not something mysterious or magic, just science.

Now, can we see conscious experience emerging from brain activity in the same way? The existence of neurons doing their stuff does certainly generate patterns of behaviour in humans and animals that can be described in language that does not apply to neurons. As it would be impossibly complicated and time-consuming to describe that behaviour in terms of neurons, we use a 'higher-level' language. This is often the language of folk psychology. We say that, for example, the woman saw a face on TV, remembered it was Boris Johnson, and became so angry that she threw a brick at the screen to vent her frustration at not being able to punch him personally. We should be using what Daniel Dennett calls the 'intentional stance': ascribing memories, feelings and intentions to the woman to explain her behaviour – concepts that could not be applied to neurons. That level of description works very well, and not only for human beings and animals. The sunflower spends all day turning on its stem, so that it is always facing the sun; that is its intention, we might say. Sometimes (perhaps not so much nowadays) the car doesn't 'want' to start on a cold morning. When I accidentally type the non-word 'consicous' on my laptop, the machine knows that there is no such word and, in its eagerness to stop me making a mistake, corrects it to 'conscious', believing that is the word I almost certainly meant. The machine has knowledge, intentions, beliefs – just like you and me. Or we might use the language of input stimulus and behavioural output, to give a functional analysis of the same activities. We might say that the mental and physical activities described in either of these ways emerge from, or are supervenient on, the behaviour of the micro-constituents of our minds or of the computer, sunflower and car.

So what's the problem? When we ascribe feelings to the car and the computer, we know – probably – that we are using anthropomorphic metaphors. They are merely easy ways of

understanding what they're doing for those of us ignorant of the innards of cars and computers. But when we ascribe feelings to people, they are not convenient ways of describing their behaviour (unless you are an unreconstructed behaviourist), but direct references to actual things going in our minds, feelings. What's more, these references are not to something observable – like behaviour seen in the street or brain activity seen on the scanner – but to something that is being experienced from the inside and cannot, even in principle, be observed from the outside. As someone who feels wonder and anger myself, I infer from the outward signs that others do too. But there seems to be no explanation for the existence of these feelings; all the functions of mental life and behaviour could go on perfectly well – as most of them in fact do – without anyone being conscious of them, just as they presumably do in machines.

This is a good moment to introduce the 'philosophical zombie', Chalmers's idea of a replica of a human being that is not conscious but whose lack of consciousness is impossible to detect. It is not a popular idea, but it has to be faced. In fact, several years before Chalmers was born, the idea occurred to *me* as a child. For all I knew, I mused, there might be only two conscious beings in the universe: one was me, the other some kind of all-powerful god who made all these other people around the place *seem* to be conscious like me when they weren't really. How could I ever be absolutely sure, I wondered. How can I now? Watching you and examining your brains would be pointless; of course the evil demon would have made you look and behave just like me. Your observable behaviour and brain states would fit the same kind of roughly rational pattern that mine do. To use your similarity to me as evidence that you were conscious would be to miss the whole point of the question! In fact, I take your consciousness on trust, seeing no compelling reason why an evil demon would want to deceive me in that way; but, however much I observed of your brains and behaviour, however much neuroscience I knew, I could never find anything to *disprove* my childhood hypothesis.

The corollary of that is another question: how would we ever know whether a very clever machine was conscious? It is a hugely important question, since, if it were conscious, we should have to consider whether we were causing it pleasure or pain; to hurt a conscious machine would be just as immoral as hurting a human being.

So I think that emergence has to go the same way as supervenience. Given all the microphysical facts about eggs, flour etc, it is impossible that those facts could hold without sponginess being instantiated – to adapt the words of Chalmers, who was talking about the emergence of liquidity from H_20. Consciousness, by contrast, if it does emerge from brain activity, is emergent in a different sense; it is a kind of brute fact emergence that does not follow inexorably from brain activity, what Galen Strawson calls 'radical' emergence – not an unexpected observable property emerging from other observable properties (which you might claim mental states in general are, including those of machines), but the property of being experienced from the inside. There is, in Nagel's well-worn phrase, *something it is like* to be me being angry; there is probably nothing it is like to be a laptop wondering which word that idiot meant to type. Brute fact emergence is simply an idea that doesn't explain anything.

6. Conclusion. So, where does that leave us? The *subjective* aspect of some mental and behavioural activity, at least of mine, is undeniable; but – to revisit the question I asked at the beginning of how it is connected to observable brain activity – it seems that it is not *caused* by the simultaneous physical activity that accompanies it, nor is it *identical* with it, i.e. simply the same thing put in different words, nor is it *logically subvenient* on the physical, nor does it *emerge naturally or logically* from the physical.

An image used by both Chalmers and Kripke is that, when God had finished creating all the physical stuff going on when one has a pain – the firing of C-fibres, or whatever – he had more work to do in order that those firings be *felt* as pain. That image may be unfortunate, in two opposite ways. It may motivate some people to espouse a kind of mental-physical dualism in order to leave room in an apparently wholly physical world for the God that they desperately want to believe in – a sort of God of the explanatory gaps; and the danger of that may motivate others to deny the hard problem of consciousness at any cost, in case it is used to justify that rôle for God. My feeling is that the idea of a supernatural being is of supreme irrelevance here. The difficulty is not a factual one of how or why consciousness arose, but how we can incorporate the subjective into the objectively observable scheme of things that science studies. How can we make the language-game appropriate to subjective experience commensurate with the language-game appropriate to science? The question is not necessarily insoluble by humans, as some believe. But, like Socrates, I feel I have heard the suggestions of various interlocutors, found them all wanting, and am consumed by a feeling – temporary, I hope – of aporia.

Recommended Reading

Chalmers, David J (1997): *The Conscious Mind: In Search of a Fundamental Theory*, Oxford

Davidson, Donald (1970): 'Mental Events', in *Essays on Actions and Events*, Oxford 2011, pp 207-227

Dennett, Daniel (1981): 'True Believers: The Intentional Strategy and Why It Works', reprinted in *Philosophy of Mind: Classical and Contemporary Readings*, ed David J. Chalmers, Oxford 2002, pp 556-568

Kim, Jaegwon (2011): *Philosophy of Mind*, Westview Press

Kripke, Saul (1981): *Naming and Necessity*, Blackwell

Owen, Adrian (2018): *Into the Grey Zone*, Faber and Faber (especially chapter 8)

Place, U T (1956): 'Is Consciousness a Brain Process?' Reprinted in *Philosophy of Mind: Classical and Contemporary Readings*, ed David J. Chalmers, Oxford 2002, pp 55-60

Strawson, Galen (2018): *Things That Bother Me*, New York Review of Books