Philosophical zombies and the problem of consciousness

Peter Coxhead

Conceivability and explanations

The general argument on which this lecture is based is the following:

> Suppose we can conceive of some state of the world which does not actually exist. Then and only then we are entitled to ask for an explanation of why this state does not exist or what causes the difference between the state we have conceived and the corresponding state which does exist.

"Conceive of" – here meaning more or less the same as "imagine" – needs to be considered more carefully than I can do in this lecture. Can I conceive of a square circle drawn on a flat piece of paper? No. Why not? Because nothing, real or imaginary, can be *described* as such a square circle. The meaning of "square" and "circle" are incompatible in this context. It's not that I don't have the ingenuity to conceive of such a thing or lack imagination; it's that whatever I conceive of or imagine cannot be correctly described as a square circle. So it's simply foolish to ask for an *explanation* of why there are no square circles (drawn on flat paper).[1]

Can I conceive of the Earth in an orbit outside the 'habitable zone'? Yes. There's nothing linguistically or logically inconsistent about this conception. Hence I can sensibly ask why the Earth has an orbit within the habitable zone. Nothing guarantees that the answer can be obtained with our current state of knowledge or that if obtained it will be of any great interest or value. All that the general argument above establishes is that there is something to be explained. In the case of the Earth's orbit, it seems unlikely that there is any deep principle which explains it; it's almost certainly the consequence of a large number of historical happenings in the formation of the solar system.

Other questions have more interesting answers. When light is shone on an appropriate material (such as that of which photoelectric cells are made), an electric current is generated. It is conceivable that the electrical energy produced would be directly proportional to the intensity of the light; there's nothing contradictory in conceiving this (and indeed this is what physicists at one time expected). Experiments show that the electrical energy produced also depends on the frequency of the light. So we can ask for an explanation of why this is the case. Providing such an explanation led Einstein to the understanding that light consisted of photons and earned him a Nobel Prize.

The hard problem of consciousness and zombies

The so-called 'hard problem of consciousness' asks for an explanation of how the properties and behaviour of a physical device – the 'meat machine' which comprises a human brain – result in subjective experiences which are *taken to be* non-physical in the sense that they cannot be explained by the laws and principles which apply to physical objects. In Chalmer's words:

> Why is it that when our cognitive systems engage in visual and auditory information-processing, we have visual or auditory experience: the quality of deep blue, the sensation of middle C? ... Why should physical processing give rise to a rich inner life at all? – Chalmers, David (1995), "Facing Up to the Problem of Consciousness", *Journal of Consciousness Studies* **2**(3), pp. 200–219

---

[1] We can conceive of *some properties* of square circles. As Aaron Sloman has pointed out, we can conceive of a thin square circle viewed sideways (because squares and circles look the same when viewed in this way). We can (perhaps) imagine projecting an image of a square onto a strangely shaped surface which makes the image circular. However, we can't conceive of figures drawn on flat paper which have *all* the properties of circles and squares because some of these properties are logically inconsistent, such as having straight sides meeting at angles and not having straight sides at all.

One way of re-casting this question is to look for a version of the general argument above which shows that an explanation is in principle possible; that there is indeed a question to be asked. The concept of a **philosophical zombie** appears to provide the right formulation.

> Suppose we can conceive of an entity which in every other way – composition, behaviour, etc. – is the same as an actual person, but does not have any subjective experiences, does not experience 'blueness' or feel pain. Such an entity is a 'philosophical zombie'. Then we are entitled to ask for an explanation of how it is that an actual person does have subjective experiences; an explanation of what it is that creates the difference between such a zombie and an actual person.

Zombies

We need to be very careful indeed in conceiving of a philosophical zombie (which I'll usually call just a zombie). The difference between a zombie and a person is easily *mis*-conceived. Kirk (in several publications) uses the idea of a 'zombie twin', but 'twin' isn't quite the right term: twins are at most only genetically identical. A better analogy comes from *Star Trek* where a transporter supposedly 'beams' a person, destroying the original and creating an exact atom-for-atom replica. Suppose the original were not destroyed, so that the transporter created a 'duplicate'. The duplicate is, at the moment of creation, identical in every respect to the original (although this will quickly change). *A 'philosophical zombie' is an exact duplicate of a person with the sole exception that the zombie duplicate does not have non-physical subjective experiences.*

There are two common errors in understanding this conception.

- *Zombies are not soulless or mindless persons.* One model of humans is the dualist one: human persons are composed of a material physical body, including the brain, and an immaterial non-physical mind or soul. Those who conceive of a person in this way could conceive of a zombie as such a person minus the mind or soul (or part of it).[1] However, this amounts to much more than the difference between a person and the kind of zombie meant here. A immaterial mind has processing ability (using computational language) – it can make choices and take decisions and can affect the body. For some dualists, an immaterial soul animates a person; for them, a soulless person could perhaps be akin to a chimpanzee but would not be capable of full human rationality. It would be much less than a zombie.

  By contrast the *only* difference between me and my zombie duplicate is that he doesn't have non-physical subjective experiences. There is no way of telling the difference between us (just as we can't tell the difference between *Star Trek* characters before and after being 'beamed up'.) My zombie duplicate is just as human in his behaviour as I am. He can make moral choices (supposing that I can), tell red from green (as he has normal colour vision like me), and respond to pain by grimacing, withdrawing, swearing, screaming or whatever. However, he doesn't experience redness or greenness or have a subjective feeling of pain. He just behaves in response to these stimuli exactly as I do (including being able to talk about subjective experiences).

- *Zombies don't lack an observer of the 'Cartesian stage'.* A common way of thinking about the subjective experiences that humans have is that there's an 'inner person', an 'ego', a 'real me', that is somehow inside me and that observes or experiences the results of brain processing – the brain has a Cartesian stage on which experiences are played out for this inner observer. It's very easy to slip into this way of thinking. For example, the image on the retina is physically upside down. It's tempting to say that the brain 'turns it the right way up', but this implies that there's someone or something inside the brain which looks at the resulting 'corrected image'. In the argument presented here, neither I nor my zombie duplicate has a Cartesian stage. Our brains process incoming stimuli, full

---

[1] There are science fiction stories in which the characters are concerned that *Star Trek*-like matter transporters will create people without souls.

stop. In my case, this processing is sometimes associated with subjective sensations; in his case, it never is.

Are zombies conceivable?

Are zombies conceivable? Many philosophers have thought so, while taking very different perspectives on how the resulting need for an explanation should be met. It's always rather dangerous to summarize complex philosophical discussions, but three positions can perhaps be identified:

- The explanation will involve concepts or entities outside those of physics. The difference between me and my zombie duplicate is a fundamental feature of the way the universe is; thus it has energy, matter, space, time, etc. and also consciousness. (E.g. Chalmers)
- Although in principle an explanation is possible, we won't ever be able to create or understand it. There's an inherent reflexivity problem; a conscious system cannot comprehend an account of its consciousness. (E.g. McGinn, ?Nagel)
- The explanation is actually quite simple, it only looks mysterious. (E.g. Dennett, although his position on zombies is not entirely clear to me.)

Others have argued that zombies are not actually conceivable, regardless of the number of people who have thought that they are. At the risk again of over-simplifying, there are three broad lines of attack (not necessarily independent):

- There is at least one logical inconsistency in any description of the concept of a zombie (like that of a square circle).
- What people actually conceive of when they claim to conceive of a zombie is not a philosophical zombie; such a conception is impossible. This seems to me to be Robert Kirk's position as explained below.
- Subjective experiences are not expressible in the kind of language required for a formal argument, whether via zombies or any other route; statements about subjective experiences are formally vacuous (although capable of being expressive).

Kirk's attack

Although at one time Robert Kirk was a 'friend of zombies' (his term for those who believe that zombies are conceivable), he has since changed his opinion. He has given various accounts of his objections; one is set out at Kirk, Robert (1999), "The Inaugural Address: Why There Couldn't Be Zombies", *Aristotelian Society Supplementary Volume* **73**(1), pp. 1-16, obtainable from http://www.psych.utoronto.ca/users/spa/reading/papers/Consciousness-Kirk.pdf.

The following is based on what I take to be the core of Kirk's argument (it differs slightly from his version), applied to a particular more concrete example.

1 Suppose my zombie duplicate and I are both looking at two objects: one red, one green. [Given]

2 Both of us can tell the difference between them; we can both obey instructions such as 'Pick up the red object.' [Definition of zombie]

3 Only I have the subjective experiences of redness and greenness. [Definition of zombie]

4 Both of us can talk about the subjective experiences of redness and greenness, but my zombie duplicate doesn't actually have these experiences, he merely behaves as if he does. [Definition of zombie]

5 I can tell the difference between redness and greenness. He can't, because he doesn't experience redness and greenness. [Given + 3]

We need to stop here and be very clear about (5), where I am taking it as given that I can tell the difference between redness and greenness. In the situation I've described, an outside observer cannot tell the difference between being able to distinguish between red and green and being able to distinguish between redness and greenness, because red and redness and green and greenness are perfectly correlated.

I can also produce a sensation of redness other than by looking at a red object (e.g. if I press on my eyeball in a certain way) and I can report on this. Of course, my zombie duplicate can also report that he has a sensation of redness under precisely these circumstances, although he hasn't, he only behaves as if he has (presumably because the process affects his physical 'red detection system'). Again an outside observer cannot tell the difference between us: in both cases some 'abnormal' process stimulates our physical red detection systems, but I correctly report having a sensation of redness, he incorrectly reports it.

So to be more precise, there's a perfect correlation between the activation of my red detection system and my having an associated sensation of redness. My zombie duplicate only has his red detection system activated; he has no sensation of redness.

6 The part of me that can tell the difference between redness and greenness is also possessed by my zombie duplicate. This is because the *only* difference between us is, by definition, the possession or not of the subjective sensations of redness and greenness. His 'missing bit' is *not* (as noted above) a mind or part of a mind which can process non-physical information like subjective sensations, but merely the sensations themselves. So if I can tell the difference between redness and greenness, so would he be able to, if only he experienced redness or greenness. [5 + definition of zombie]

Again we need to stop. Point (6) is very odd indeed. My zombie duplicate is totally physical; he doesn't have any non-physical subjective sensations. So how can he have some physical system which could distinguish redness from greenness if only he experienced them, given that redness and greenness are (by the definition of a philosophical zombie) non-physical subjective phenomena? How could there be such physical devices as 'redness distinguishers' (given that these are different from 'red distinguishers')? And yet he must have such abilities, because by definition all I have to do to turn him into a precise replica of me is to add sensations, not add sensations plus the ability to process sensations.

I think we can now go in one of two directions; Kirk chooses (7a).

7a Point (6) shows that we cannot conceive of my zombie duplicate, who would have to possess a physical system which could distinguish between the non-physical subjective sensations which he doesn't have. Such a zombie is not actually conceivable. When we think we are conceiving of a zombie, we are imagining something different: an entity which not only does not have subjective sensations but also does not have the apparatus required to recognize or distinguish such sensations. But such an entity is something like a partially mindless or soulless person, not a philosophical zombie.

7b The 'given' part of (5) is wrong. Although I think I can recognize and distinguish between redness and greenness, I cannot. I am actually always distinguishing between red and green. In some sense there is no such thing as redness as distinct from red; the idea that there is such a thing is an illusion. There is thus either no difference between me and my zombie duplicate or such a small difference that it is of no importance. This seems to me to be more-or-less the Dennett approach (but I find Dennett rhetorically interesting but logically hard to follow).

Conclusion

I believe that:

- Dualism (where the supposedly non-material part of a person has processing power) is capable of being expressed in a self-consistent manner, but has been shown to be wrong (going back at least as far as Ryle's *Concept of Mind*).

- Zombie-ism (where the supposedly non-material part of a person is nothing but sensations) is logically incoherent.

Assuming that there are no other accounts to be considered (but see Addendum 2), it follows that *subjective sensations are not somehow outside the realm of material or physical systems*, since the accounts that suppose that they are non-material or non-physical are either incorrect or incoherent.

I am aware that this argument seems unsatisfactory to those who are 'bugged' by how material systems result in sensations. For example, Susan Blackmore (http://www.susanblackmore.co.uk) gave a seminar in the School last year and both there and in discussion afterwards made clear how this issue really 'gets to her'. It doesn't get to me, so I'm perhaps not the best person to discuss why the argument above feels unsatisfactory. However, in general, negative explanations, like *reductio ad absurdum* arguments, feel less satisfactory to most people than positive ones. At present I'm not convinced that there are any clear and satisfactory positive explanations of the relationship between physical devices such as brains and subjective sensations.

Addendum 1

There is another, wider argument against zombies, which I find attractive, but which in my experience is even less satisfactory to most people.

In brief outline, this argument holds that any statement supposedly referring to subjective sensations is either referring to the objective properties of things or is merely expressive not descriptive (so is not a statement). On this view, although *redness* is grammatically a noun, it is not the name of anything, any more than *ow!* is the name of a sensation of pain. Under normal circumstances, asserting "I have a sensation of redness" means nothing more than (the literal meaning of) "I see red" (under abnormal circumstances, such as pressure on the eyeball, it means something like "I see what appears to be red"). There's no way, on this view, of making those statements about sensations which are needed in order to establish a difference between me and my zombie duplicate. All supposed statements about this difference are not actually statements. "Whereof one cannot speak, thereof one must be silent" (Wittgenstein, *Tractatus Logico-Philosophicus*). Unfortunately, philosophers have not heeded Wittgenstein's advice.

It's partly for this reason that I have deliberately not used the term *quale* (plural *qualia*). I think this language can produces a misleading appearance of 'objectivity'. The 'redness quale' seems somehow more solid than 'my sensation of redness'.

Addendum 2

As Aaron Sloman has clearly pointed out, the behaviour of a computer program is different from that of a 'typical' physical or material system in some interesting ways. Consider two computers which have different chips and different operating systems but are both running the same Java program via each computer's version of the JVM. Then at the lowest physical level – connectivity, electrical signals, etc. – the two computers have very different behaviours. However, at a higher level – the functionality of the Java program – they have the same behaviours, give or take a few small differences in visual appearance.

This leads to the possibility of other kinds of zombie. In terms of William Gibson's 'cyberspace' (dating back to 1982) or *The Matrix* series of films, we can conceive of a 'cyberspace zombie'. It may be conceivable that although I have subjective sensations when the 'program that is me' is running on my brain, when it is run in cyberspace, on different hardware, it does not have such sensations, but merely appears to do so. Is such a 'cyberspace zombie' conceivable? If so, does it tell us anything about the relationship between the 'program that is me' and my sensations? If not, why not?