# Thinking About Zombies

*Paul Sperring*

## *1. Zombies*

A number of contemporary philosophers have made use of the idea of zombies to defend a certain position within the philosophy of mind. What they have in mind are creatures very distinct from the sorts of zombies imagined by horror-film makers or Haitian occultists. *Philosophical* zombies are just like you and me in many respects. In fact, they are exactly like you and me in all but one important respect. They are physically, behaviourally and functionally identical to us but utterly without any conscious states.

Let us imagine that I have a zombie counterpart. We will call him $Paul_z$. Let us further imagine that this counterpart inhabits a world pretty much like the actual world. We'll call it $world_z$ (or $W_z$). This counterpart will, in $W_z$, do and say whatever I do and say here in the world. Let us imagine, then, that $Paul_z$ is currently tapping away at a word-processor writing a paper about zombies in $W_z$. If we were witness to $Paul_z$'s actions then we would be able to discern no difference between what he is doing there and what I am doing here. All goings on in $W_z$, from the outside, would appear to be identical to the goings on in the actual world.

My experiences, tapping away at the keys, will have features, however, that $Paul_z$'s will lack. Properly speaking, $Paul_z$ will lack *phenomenal* experiences entirely. I am currently having a variety of phenomenal experiences, each with a distinctive qualitative feel: the texture of the keys under my fingers; the tapping and clicking noises made by my fingers striking the keys; the varied colour experiences of the computer monitor and objects surrounding it on my desk; the smell and taste of the tea just now sipped. All of these things will be denied to $Paul_z$ in his phenomenally textureless, noiseless, colourless, odourless and tasteless $W_z$. Of course $Paul_z$ will appear to have all of these experiences, and when someone asks about his cup of tea he will say 'very nice, thanks, just right, not too hot', or something like it.

It will appear, from the outside, that $Paul_z$ is having all sorts of conscious experiences. An observer, on being told that $Paul_z$ had no conscious experiences at all, might, of course, be puzzled about how it was that $Paul_z$ could type meaningful sentences employing references to phenomenal experiences unless he had had such experiences. However, we will put this worry aside here. Let us merely reiterate that the zombie counterpart has no phenomenal experiences at all, despite appearances.

## *2. Attacking the Identity Thesis*

So, this is what the philosopher has in mind when she introduces the idea of a zombie into her discourse. What purpose is served by thinking about zombies? They are introduced, usually, as a means of showing that the identity thesis about mind and body is false. The argument runs as follows:

1. If *a* is identical to *b* then *a* is necessarily identical to *b*.

2. If *a* and *b* are identical then there is no possible world where there is *a* but not *b*.

3. There is a possible world where there is *a* but not *b*.

4. Therefore, *a* is not identical to *b*.

To cash this out we can see that in $W_z$ there are functioning brain states of a certain sort[1] (cases of *a*) but no conscious states (cases of *b*), as with $Paul_z$. But if conscious states were identical to brain states, hence necessarily identical, then there couldn't be a world where there were functioning brain states of a certain sort without conscious states. But we know that it is conceivable that there is a world where brain states occur without conscious states – we have just conceived of the world above ($W_z$), where Paulz is typing – and since whatever is conceivable is possible it must follow that brain states and conscious states are not identical.

To make the argument transparent let us run it again, attempting to identify particular brain states (Z-fibre firings[2]) with particular conscious states (phenomenal experiences):

1. If Z-fibre firings are identical to phenomenal experiences then Z-fibre firings are necessarily identical to phenomenal experiences.

2. If Z-fibre firings and phenomenal experiences are identical then there is no possible world where there are Z-fibre firings and no phenomenal experiences.

3. In $W_z$ there are Z-fibre firings but no phenomenal experiences.

4. Therefore, Z-fibre firings are not identical to phenomenal experiences.

If this argument is sound then the identity thesis about mind and body is in trouble. Is it sound? I want to suggest a couple of lines of attack against the argument, but in its strongest versions it has been taken by a number of philosophers to be compelling.[3] One way of undermining the argument would be to deny that zombies are conceivable. If it could be shown that we cannot actually conceive of such things then we lack the grounds for claiming that they are possible - that is, $W_z$ would not be a possible world. This would be an interesting strategy to adopt, but the defender of the zombie argument might be puzzled as to what it is that they have been thinking about all along. The objector will have to explain the apparent ease with which we think about zombies, that is, if the objection is that zombie thoughts are impossible thoughts then what is it that we have been conceiving of if not zombies? Ersatz zombies perhaps? I will say a little bit about this later, but first I want to explore another possible way of attacking the argument from zombie conceivability, *viz*., the denial that conceivability entails possibility.

## *3. Does Conceivability Entail Possibility?*

Despite its appeal many commentators have objected to the claim that whatever is conceivable is possible. The claim, and a forebear of something like the zombie

argument, can be traced right back to Descartes who also helps himself to a refutation of materialism on the strength of it. Descartes' argument depends on the following principle.

(C⇒P) If we can conceive of some state of affairs S then S is possible.

There is *prima facie* plausibility to this claim. It can easily be shown how conceivability is a pretty good guide to what is and isn't possible with some examples. We can conceive of Brighton and Hove Albion winning the FA cup this year, or we can conceive of a British rail company running all of its trains on time all of the time. Someone might object that these are unlikely scenarios, but it would be rather odd for that someone to claim that they are *not possible* states of affairs. On the other hand we cannot conceive of a square circle or a married bachelor, and we are evidently right to conclude that these things are not possible states of affairs. These latter are inconceivable for Descartes because they are self-contradictory, we simply cannot, for instance, think the thought of a person who, simultaneously, both has and does not have the property of being married, whereas there is nothing self-contradictory in thoughts of the former type involving football matches or efficient public transport systems. So conceivable states of affairs are states of affairs that are possible (or *not im*possible, which is the same thing). So if Descartes is right here then if we can conceive of a certain something then that thing is possible.[4]

It is on this basis that Descartes constructs his argument for dualism. He thinks that we can conceive of minds without bodies. The parallel with the zombie argument will be obvious, but instead of mindless bodies we get disembodied minds (ghosts, rather than zombies). The following is Descartes' argument as it appears in Meditation VI:

> First, I know that everything which I clearly and distinctly understand is capable of being created by God so as to correspond exactly with my understanding of it. Hence the fact that I can clearly and distinctly understand one thing apart from another is enough to make me certain that the two things are distinct, since they are capable of being separated, at least by God. The question of what kind of power is required to bring about such a separation does not affect the judgement that the two things are distinct. Thus, simply by knowing that I exist and seeing at the same time that absolutely nothing else belongs to my nature or essence except that I am a thinking thing, I can infer correctly that my essence consists solely in the fact that I am a thinking thing. It is true that I have […] a body that is very closely joined to me. But nevertheless, on the one hand I have a clear and distinct idea of myself, in so far as I am simply a thinking, non-extended thing; and the other hand I have a distinct idea of body, in so far as this is simply an extended, non-thinking thing. And accordingly, it is certain that I am really distinct from my body, and can exist without it.[5]

To repeat, Descartes is accepting the claim here that whatever is conceivable is possible, but he is not appealing, as in the zombie argument, to possible worlds. Instead he employs God to make the conceivability entails possibility thought (hereafter C⇒P) transparent. For any clear conception that we have of a state of affairs S God could make S obtain.[6] So if I have a clear and distinct idea of something then that thing just is possible. Once we accept this then we can see how the

conclusion follows. If mind and body were identical then a clear and distinct conception of the essence of one would necessarily bring with it a conception of the other. But if we can conceive of states of affairs where there are minds but nothing physical or bodies but nothing mental then mind and body are possibly distinct. According to Descartes we can have a conception of mind that excludes any physical features and a conception of body that excludes any mental features so they cannot be identical. Hence dualism.

However, it might be objected, just because I can conceive of minds and bodies as separable does it really entail that they are possibly separable? Does conceivability in this instance entail possibility?

Antoine Arnauld constructed a counterexample that questioned Descartes' use of the C⇒P principle.[7] In Arnauld's objection he asked us to imagine a confused geometer who understands some basic properties of triangles but who is not familiar with all of the properties of them. In particular the geometer does not know that all right-angled triangles have a certain property ϕ (the property expressed by Pythagoras' theorem). He then conceives of a right triangle T that lacks ϕ, and (by C⇒P) concludes that it is possible that there is some such object (T without ϕ). However, T, if a right triangle, *necessarily* has property ϕ, so it is not possible that T without ϕ obtains. But if conceivability does not in all instances entail possibility then need we accept that the conceivability of disembodied minds entails that there could be such things?

Descartes, in answering Arnauld, points out that unlike the confused geometer he, Descartes, has a *clear and distinct* conception of mind and body. The geometer concludes that T without is possible precisely because he doesn't understand what essentially belongs to T and all right triangles. If indeed we supposed him to be an even worse geometer then there is no telling what sorts of things might, by his own lights, be included in the space of geometric possibilities – maybe triangles whose internal angles were greater than 180º, or perhaps even square circles. So it is not conceivability *simpliciter* that entails possibility but clear and distinct conceivability.[8] Thus we have a stronger version of C⇒P.

($C_{cd}$⇒P) If we can clearly and distinctly conceive of S then S is possible.

What does Descartes mean exactly by 'clear and distinct' conceivability? In Meditation III he spelt out its use as a criterion for certainty – anything that can be so conceived must be true. Thus anything that cannot be doubted, that must be assented to when thought about (the clear part), independent of elements that are doubtable (the distinct part), could not possibly be false. The *cogito* shows this clearly: I cannot doubt that I exist, it is impossible to think that I do not (I *clearly* conceive of myself, as a thinking thing, as existing), and I am thinking here only of my capacity to think, independent of (*distinct* from) thoughts about my body or some other thing.

So, having realised that he cannot but be certain about his own existence Descartes asserts that there must be something in the manner of his knowing this that can be generalised as a test for certainty. Anything else known in this manner will have the stamp of certainty.

There is, however, an immediate note of caution struck in Meditation III. Descartes accepts that some things might at first appear to be known with clarity and distinctness, but turn out, on more careful reflection, not to be. In a pre-philosophical state this is how things are with most of us – e.g. when we ordinarily assume that our senses give us reliable and accurate knowledge of the world as it really is. Having recognised that it is a problem that needs to be addressed, however, it is not at all clear that Descartes has a solution to it. And it is exactly this problem that gives Arnauld's objection its teeth.

Arnauld's attack is supposed to be defeated by $C_{cd} \Rightarrow P$ because the confused geometer had no clear and distinct conceptual grasp of triangles and their necessary properties. He had an unclear, imprecise conception of what belonged to triangles and hence had no warrant for the claim that what he thought about them was genuinely possible. Now, Descartes' stronger version of $C \Rightarrow P$ might work if we could pin down (a) what we mean by 'clear and distinct' conception and (b) when we have epistemic warrant for claiming that the conception under consideration counts as a veridical case.

Unfortunately for Descartes he is unable to provide a satisfactory account of either. In the case of (a) – aside from appealing to some unsatisfactory characterisations, such as 'that which is manifest to the natural light of reason' – Descartes can only appeal to cases where we do grasp the thing in question clearly and distinctly as *illustrative* of the conditions required. The best (and perhaps only) illustrative example, as we saw above, is the *cogito* where we are just compelled to see its truth as soon as it is presented to us.[9]

With respect to (b) this just brings us right back to Arnauld's objection. It seems right to say that the geometer *would* change his mind about thinking the triangle in question possible if his conception was much clearer. But what if the geometer in question had thought his conception to be clear and distinct. It does not seem enough to say that he was just wrong to think he conceived things clearly and distinctly – intuitively, it is obvious that he just does not see things aright – because Arnauld is looking for a warrant independent of the *seeming to be* in the grip of clarity and distinctness. In other words, for our discussion, what allows us to distinguish cases of *apparent* $C_{cd} \Rightarrow P$ and cases of g*enuine* $C_{cd} \Rightarrow P$? This goes right to the core of the question whether (and when) conceivability entails possibility. If we have no epistemic warrant then we are merely appealing to the gut, and if this was not enough for acceptance of $C \Rightarrow P$ then it is difficult to see how it will suffice for $C_{cd} \Rightarrow P$. What one needs from a refined version of $C \Rightarrow P$ then is some sort of internal guarantee that allows one to hold one's conceptions up to the light and see that they are genuine cases of conceivability that entail genuine possibilities. That is, we need something that will ensure that conceivability infallibly picks out possibilities.

Before considering whether there are any better versions of $C \Rightarrow P$ that will do the work here I want to see if something like Arnauld's objection might be deployed against the zombie conceivability argument. Might the person who conceives of the zombie be like the confused geometer? It might be thought that the two cases are disanalogous since the geometer was thinking of something that was logically impossible whereas there seem to be no logical or conceptual constraints on our thoughts about zombies. Better thinking *a priori* on the geometer's part would have revealed to him why there could not be a right triangle that lacked the property. Could

better thinking *a priori* rule out zombies? What if we, committed to the truth of the identity thesis (perhaps having independent reasons for accepting it), just stipulated that if anything had property *a* it would of necessity have property *b* – and then it would be an analytic truth that there could be no *a*'s without *b*'s, i.e., no zombies, so they can be ruled out *a priori*. We could simply claim in advance that if pains simply were Z-fibre firings then just as there could not be a world where one had Z-fibre firings and no Z-fibre firings – $\Box\sim(P \& \sim P)$[10] – then, on the assumption of identity, there could not be a world with pains and no Z-fibre firings. Or to put the thought another way, if consciousness were just a physical/functional *concept* then there would be a conceptual bar to thoughts about zombies.

The reason that this could be rejected as an option is not merely that we do we not know that the identity thesis is true yet, but because it seems intuitively right to say that no matter how much evidence that we might accrue about the brain, no matter if we had a complete physics in place, there would still be no *a priori* warrant for ruling out the separability of mental and physical states. And if we cannot rule out a claim *a priori* then it follows that there is no conceptual problem holding it to be the case. I think that there is much more that needs to be said about this – for my own part I suspect that there could be some deep connection between physical states of affairs and phenomenal states that might, on further discoveries, rule out *a priori* the zombie claim. For now, however, we will allow that zombies are conceivable. The question remains then whether they are, accordingly, possible, and I think that the Arnauld objection does at least suggest that a gap might be opened up between conceivability and possibility. Can that gap be plugged?

## *4. Ideal Conception*

Can we refine $C \Rightarrow P$ so that it delivers infallibly? Some philosophers have appealed to ideal forms of conceivability as illuminating genuine possibilities. David Chalmers[11], for instance, distinguishes *prima facie* from ideal conceivability, where someone conceiving in the former sense wouldn't have sufficient warrant for the claim that such and such was possible, but someone who conceived of things in the latter sense would. So in Arnauld's confused geometer we clearly had a case of *prima facie* conceivability. Any ideal conceiver would not have made the elementary error of thinking that the triangle conceived of could lack the salient property. Ideal conceivers need, at least, to be experts in the areas to which the conceivings apply. However, this looks just a bit too like Descartes' appeal to clarity and distinctness in our concepts, and wouldn't deal with the following case, which Chalmers mentions. Frege thought that there was a set of all sets, presumably having thought carefully about the matter. Later, however, Russell came along and showed that there couldn't be such a thing because the very idea of a set of all sets generates a paradox. Now Frege, unlike the confused geometer, was an expert in the field in which he had his conceivings.

One might argue in fact that, at the time Frege thought that there was such an object, hardly anyone was better placed to adjudge whether there could be a set of all sets. But, after Russell, we know that such a thing is an impossible object. Here is a case of conceivability failing to reveal possibility, but not, it seems right to say, mere *prima facie* conceivability. This was, however, a case of *secunda facie* conceivability, according to Chalmers, which is, as its name suggests, a step up from our first, ill-

considered, conceivings, but falls short of ideal conceivability. Russell, however, was in the position of the ideal conceiver. So what, exactly does it mean to conceive of something ideally? The following might be a first shot at framing such a principle.

($C_{irr}\Rightarrow P$) If S is conceivable on ideal rational reflection then S is possible.

But what is meant by *ideal rational reflection?* The thought seems to be that a statement S would be ideally conceivable 'if an ideal reasoner would find that it passed the relevant tests'[12], i.e., tests such as attempting to rule S out *a priori*. Chalmers himself doesn't find this entirely satisfactory because of the difficulties faced in making sense of the notion of an ideal reasoner, that is, whether for any imagined 'ideal' reasoner we could imagine one who is even smarter (more ideal), so he instead he appeals to the notion of 'undefeatability by better reasoning.' So we get something like.

($C_{\sim rd}\Rightarrow P$) If the justification for S cannot be rationally defeated then S is possible.

On this version we can see clearly that Frege lacked final warrant for his belief in the set of all sets because better reasoning (Russell's) would have defeated the belief. But this again begs the question when are we in a position to say that we have everything that we need to satisfy $C_{\sim rd}\Rightarrow P$? I just can't see how this looks in better shape than $C_{cd}\Rightarrow P$.

We believe things when we don't have before us defeaters for our beliefs, so in a sense we get this for free. I think that such and such is conceivable until I am given grounds for not taking it to be so (someone introduces a defeater such as a clear counterexample to my $C\Rightarrow P$ claim). What we have to work much harder for (and I am pessimistic about reaching this goal) is a justification that we are in the position where no further defeaters are possibly forthcoming - and this just reiterates the question asked all along, 'when are our conceivings infallible guides to possibility?'

So if it looks like there is always going to be a gap between what we conceive and what is possibly the case then does this give us reason to suppose that even were zombies conceivable they needn't be possible? I think so. And if there is always doubt that conceivability hooks onto real possibility then at the very least we can assert that the case against materialism has not been conclusively established on the basis of zombie conceivability alone.

Paul Sperring
*Richmond-upon-Thames College*

[1] Obviously there could be functioning brain states that occurred without any conscious states at all. There are no conscious states that accompany the brain states that are involved in the regulation of my breathing, for instance. I have in mind just the salient brain states that are accompanied, in non-zombie cases, by phenomenal states (in the literature, C-fibres firing accompanied by feelings of pain).

[2] By 'Z-fibres' I mean just whatever particular brain states are identified, by the neuroscientist, with the phenomenal experiences had by the possessor of those brain states.
[3] The stoutest defence of the argument for dualism from zombie conceivability can be found in David Chalmers' book *The Conscious Mind* (New York: Oxford University Press, 1996).
[4] Which isn't to say that what is possible is just defined as what we think about coherently. Possible states of affairs exist independently of our capacity to conceive of them, which Descartes illustrates by appealing to, for instance, complex geometric properties. These properties exist whether I choose to think of them or not, and I am not free not to think about them other than as they are once they are discovered to be essential to the object in question. Possibility, for Descartes, is fixed by metaphysical not epistemological limits.
[5] *The Philosophical Writings of Descartes, Volume II* (Cambridge: Cambridge University Press, 1984), p.54
[6] It might be objected that if the argument hinges on the truth of theism then it has little chance of going through. However God is not necessary for the argument, Descartes is merely employing God as a device to delimit the space of possibilities (much in the way that modern arguments employ possible worlds talk). So if a certain state of affairs is conceivable then (by whatever power) that state of affairs is possibly instantiable exactly as conceived.
[7] *The Philosophical Writings of Descartes, Volume II*, pp.139-143
[88] Or one might say, that it is not *apparent* cases of conceivability that entail possibility but *genuine* cases, and Descartes clarity and distinctness criterion provides us with a means of distinguishing between the two. This claim is not without its problems as we shall soon see.
[9] There is something in this – we do seem to have unimpeachable warrant for the claim that our existence is indubitable – but whether other claims can be known with quite the clarity and distinctness of the *cogito* is doubtful.
[10] Meaning 'necessarily, nothing can be both itself and not itself.'
[11] David Chalmers, 'Does Conceivability Entail Possibility?', in Tamar Szabo Gendler and John Hawthorne (eds.), *Conceivability and Possibility* (New York: Oxford University Press, 2002), 145-200.
[12] ibid, p.148.