# Against Semantic Externalism and Zombies

DRAFT

Paul Tappenden paulpagetappenden@gmail.com

31 October 2016

**Abstract**

It is widely believed that the semantic contents of some linguistic and mental representations are determined by factors independent of a person's bodily makeup. Arguments derived from Hilary Putnam's seminal Twin Earth thought experiment have been especially influential in establishing that belief. I claim that there is a neglected version of the mind-body relation which undermines those arguments and also excludes the possibility of zombies. It has been neglected because it is counterintuitive but I show that it can nonetheless be intelligibly worked out in detail and all obvious objections met. This suggests that we may be faced with a choice between embracing a counterintuitive interpretation of the mind-body relation or accepting that a currently very promising theory in cognitive science, Prediction Error Minimization, faces a fundamental problem. Furthermore, blocking that threat entails that any physicalist/materialist theory of mind is freed from the spectre of zombie worlds. The proposal also makes the ideas of personal teleportation of mind uploading more plausible.

## 1. Against Semantic Internalism

Analytic philosophy took a Hegelian turn in the late 20th century, as Tyler Burge acknowledged, seeing Ludwig Wittgenstein as its harbinger. Burge contrasted modern semantic externalism with the 'elderly Cartesian tradition' that put 'the spotlight on what

exists or transpires "in" the individual – his secret cogitations, his innate cognitive structures, his private perceptions and introspections, his grasping of ideas, concepts or forms (1979, p.73). Burge contrasted this Cartesian 'individualistic' perspective with the 'Hegelian preoccupation with the role of social institutions in shaping the individual and the contents of his thought' (*ibid.*). Though discussion of the role of social institutions had been present in the analytic tradition's study of language and mind, what abruptly precipitated the Hegelian turn was the introduction of 'Twin Earth' thought experiments by Hilary Putnam (1975, p. 139ff.) and their further development by Burge. It is these arguments which have been most influential in establishing the widespread rejection of an unqualified semantic *internalism* in current analytic philosophy. Internalism does not deny a role to social interaction in the development of minds and language, but insists that the neural structures which are normally forged during evolutionary and social processes are autonomous in a sense which I shall explain shortly.

The internalist-externalist dichotomy in views about mentality can be couched in terms of the representational view of mind which is common in contemporary cognitive science. That will provide a useful framework for explanation. The representational view has enough currency for what follows to be worthy of attention and even if a different framework for the science of mind is adopted that may also be adaptable to the purpose in hand. Briefly, the representational view holds that mentally-driven behaviours are causally mediated by objects which bear semantic content. Explaining the behaviour of a creature requires the attribution of contents which specify the ways in which the environment and internal states of that creature are represented. The objects which function as mental representations, the vehicles of content, have traditionally been thought to be cerebral though that is disputed by advocates of the so-called extended mind idea, of which more later.

The question then is whether the semantic content of some mental representations is determined extra-cerebrally. The question is brought into sharp relief by the idea of Donald Davidson's Swampman (1987, pp. 443-4) or, better, what has come to be known in cosmological circles as a Boltzmann brain and may be more perspicuously dubbed an accidental cerebral episode (ACE). An ACE is an object which is isomorphic to a normal brain for a duration of at least a substantial fraction of a second (it is an atom-for-atom duplicate). It pops into existence by so-called quantum accident and is sustained by what Bertrand Russell would have called 'a climax of improbability' (1954, p. 33). Do *all* the structures and processes in an ACE which are isomorphic to the corresponding items that are representations in a normal brain bear the same semantic content as the normal representations? Externalists say 'no'.

An intuition apparently supporting externalism was well expressed by Putnam (1981, pp. 1-4). Imagine an ant crawling on sand which, by chance, happens to trace 'a recognizable caricature of Winston Churchill' (the tracing of Alfred Hitchcock's silhouette is brought to mind). We can agree that the mark in sand does not represent Churchill, it is a meaningless scrawl. Now imagine aliens on a treeless planet where a sheet of paper just happens to drop from the sky which seems to be a picture of a tree but is in fact 'the accidental result of some spilled paints' (*ibid.*, p. 4). If one of those aliens looks at the sheet and forms a mental image, is it an image of a tree? It can seem obvious that it is not. If a brain comes to contain something which is just like a representation of a tree but which has no causal connection with trees how can it be a representation of a tree?

However, it is not immediately clear that this argument is enough to make the externalist case. The argument points towards the claim that objects in an ACE cannot be representations of trees because, like that alien's mental image, they have no causal connection with trees. But the internalist can attempt to counter with the concept of

'plugability'. Take a normal person whose actions and speech suggest that s/he hosts mental content about trees. And take an ACE which is isomorphic to that person's brain. Now substitute the ACE for the normal brain, preserving all the right connections. The ACE will no longer need a climax of improbability for its continued existence as it has become normally embodied. We can suppose that the post-operative person will behave and speak in exactly the same way as the original. The internalist might claim that this is enough to show that the ACE's representations bear the same content as those of the original brain. What is important in thinking about trees is to have the internal causal structure which allows a person to interact successfully with trees as trees and apparently talk intelligibly about trees. After all, the internalist may argue, the very idea of mental content is a hypothesis aimed at explaining behaviours. The behaviours of a normal person and of that person when their brain has been replaced by the remnants of an ACE will be, presumably, isomorphic. So why should mental contents not be assigned to the person into whose body the ACE is grafted if those contents would be assigned to the normal person?

But Putnam initiated a series of arguments which pushed the externalist case further than those reflections on ants and trees. These are the Twin Earth arguments which derive from his seminal paper. They have presented a challenge to semantic internalism which has not yet effectively been met by general agreement. Details aside, Putnam's central idea is that if two physically isomorphic doppelgangers were to exist in physically anisomorphic environments there could be differences in the semantic content of some of the mental representations of the environments by the doppelgangers. My aim is to counter that idea more effectively than has been done to date.

Putnam's battle cry was 'Cut the pie any way you like, "meanings" just ain't in the head!' (1975, p. 144). By way of preparation for meeting that challenge in a novel way I shall

begin by considering an argument from contemporary cognitive science which supports the idea that minds are in heads.

## 2. Contemporary Mind-Brain 'Identity'

Ideas about brain function and computation have proceeded apace in recent years and in the wake of those developments Jakob Hohwy has recently defended the idea that minds are indeed in heads (2014). Hohwy's argument is based on an analysis of brain function known as Prediction Error Minimization (PEM) (2013). Whilst not attempting to address Twin Earth arguments Hohwy does refer to them, writing 'To me, it seems that PEM is exactly the kind of view targeted by the earlier metaphysical debate' (2014, p. 24, note 8).

It is easy to see why that should seem so. PEM is a representational theory of mind. It involves an inferential mechanism where content-bearing representations are in causal interaction. The nature of the causal interaction depends on the contents, so if the contents are not wholly determined within the mechanism itself it cannot function as an inferential engine. And Putnam's Twin Earth thought experiment, which was at the core of that 'earlier metaphysical debate' implies that some contents cannot be fully determined within a PEM mechanism.

Before coming to Putnam's famous example, first consider a modified version of his challenge. It is currently unknown whether electrons have a substructure or not but suppose that early in the Big Bang two kinds of electron were produced, one comprising five particles and the other seven. Call them pentacles and septacles. The two kinds were created in separate regions of the universe which have now expanded to become much larger than our local galactic clusters and conditions have nowhere been extreme enough for electrons' sub-structure to be revealed. That being so, it seems obvious that electrons in our local

environment must now, unbeknown to us, be either pentacles or septacles. Suppose that they are pentacles.

Now imagine that there is a far-off region of the universe which is exactly like ours in every way, it is physically isomorphic, except that the electrons are septacles. We have doppelgangers in that world, the only difference between us being that the electrons in their bodies are septacles, not pentacles. But it is implausible that that hypothetical sub-structure of electrons could have a causal role in our neural mechanism. It is much more plausible that the mechanism operates at the intercellular and molecular levels. So our mental mechanisms and those of our doppelgangers are type-identical. And yet when we think about the particles in cathode ray tube beams we think about pentacles and when they think likewise of old-fashioned television screens they think about septacles. To deny that is like denying that water was $H_2O$ back in 1750, before anyone knew about molecular constitutions, that's Putnam's point. And if quibbles about fuzzy water concepts are set aside and Putnam's point is well-taken the implication is that PEM requires that electrons in our environment *cannot* be either pentacles or septacles given the supposition that we have doppelgangers in distinct regions of the universe where electrons exist in those two varieties. Because if our electrons were either pentacles or septacles we and our doppelgangers would entertain different mental contents when thinking about electrons so semantic internalism would be false.

What is at issue here is not a matter of indexicality, as Burge made clear when he wrote:

> [*D*]*e re* belief attributions are fundamentally predicational. They consist in applying or relating an incompletely interpreted content-clause, an open sentence, to an object or sequence of objects, which in effect completes the interpretation. What objects these open sentences apply to may vary with context. But, according to the picture, it remains

> possible to divide off contextual or environmental elements represented in the propositional attitude attributions from more specifically mentalistic elements.
>
> (1982, p. 98)

Given the electron hypothesis, an utterance of 'that electron is a pentacle' can be understood as 'there is something x, and x is a pentacle'. Burge's thought implies that if there are two doppelgangers, one pointing to a pentacle and the other pointing to a septacle there is no need to assign different mental contents to them simply because they index different objects. The numerical difference between the objects is in the environment, not the minds of the subjects, because indexing an object simply involves assigning a value to the mental equivalent of a free variable.

This clarifies what is involved in Putnam's challenge. Given the story about electrons and our doppelgangers, the implication of the semantic internalism involved in PEM is not that electrons in our environment are either pentacles or septacles but we do not yet know which. It is rather that electrons in our environment cannot be *either* pentacles *or* septacles. Semantic internalism requires that the meaning of our term 'electron' is not in any way determined by undiscovered features of the constitution of electrons. And as with linguistic representations, so with mental representations. PEM cannot function with mental representations of electrons if the content of those representations is not determined within the representational mechanism itself. The content cannot be determined by an as yet undetected aspect of the environment. How is that to be reconciled with our intuition that if water is $H_2O$ now then it must also have been $H_2O$ even when nobody on Earth had detected its molecular constitution? Before resolving that problem it will be useful to get clear about a debate to do with PEM to which Hohwy devotes much discussion (2014: p. 10ff.).

## 3. A Mind's Extent

In Hohwy (2014) an important concern is to counter an idea dubbed 'the extended mind', Clark and Chalmers (1998). Andy Clark (2016) has since further defended the extended mind idea against Hohwy (2014), arguing that it is compatible with PEM, so that debate is currently ongoing and it will be useful to see how a resolution is *not* required in order to defend PEM against Putnam's challenge.

PEM is a theory which involves content-bearing mental representations. Hohwy defends the idea that those representations are neural and he sees the boundary which encloses them as the dorsal horn of the spinal chord (2014, p. 18). The extended mind thesis is that mental representations can be wholly or partly constituted by extra-neural objects. The argument is not about how the contents of the representations are determined but about the nature of the representations themselves.

Clark argues that part of the computational mechanism of PEM may be outside the central nervous system and even outside the body. Environmental objects can partly or wholly constitute content-bearing mental representations and an organism's interaction with its local environment can constitute part of the inferential mechanism. It is important to note that this supposed extension of the mind into the environment is local. It is only in so far as an organism can causally interact with the environment that its mind extends into it. That is different from semantic externalism which posits that it is the very constitution of the environment which can determine mental contents. If water is $H_2O$ then that is so both here on Earth and on a planet in a distant galaxy.

The contrast between Hohwy's neurocentrism and Clark's extended mind can be thought of as like a spider and its web. Hohwy claims that the mind is wholly contained within the neural spider whilst Clark contends that it extends into the web of causal

interactions between the spider and its local environment. But Putnam's challenge to PEM is independent of this debate which is just about the local compass of a mind's extent. He declared that meanings 'ain't in the head' but that could just as well be 'ain't in the web'. There is a wide world beyond the web.

As Hohwy puts it:

> It is thus a requirement on any PEM-based account that it allows, and is in principle able to describe, the boundary relative to which prediction error is being minimized, from behind which the mind tries to infer the hidden causes on the other side. Failing that, there will be no evidence for the existence of the agent in question.
>
> (2014, p. 7)

What is going to be important in meeting Putnam's challenge to PEM is to argue that there is no good reason to think that the contents of mental representations are not wholly determined within the boundary behind which an agent is confined. And central to that argument is going to be how we think about the mentality of physically isomorphic agents in environments which are anisomorphic beyond the boundary. Whether those agents are matched at the level of the central nervous system alone or are taken to include the local environment with which it interacts is an irrelevant detail and so, for the sake of simplicity, I shall assume neurocentrism in what follows. But before tackling Putnam's idea it will be helpful to consider a related issue which poses a different challenge to PEM.

## 4. The Challenge of Consciousness

There has been much work on what have come to be known as the neural correlates of consciousness (NCCs). In discussing that work, Hohwy and Tim Bayne identify two distinct types of correlation and what they call the 'NCC project', going on to write:

> In many ways it would be preferable to describe this project in terms of the search for the neural *substrates* or *basis* of consciousness, but in our view the NCC terminology is now so deeply entrenched that any attempt to displace it is bound to fail.
>
> (2015, page in book needed)

What motivates their use of the italicized terms is this:

> the disagreement about the metaphysics of consciousness has little direct bearing on the NCC project, for all that the project requires is that certain neural states 'underlie' consciousness, either by way of identity, constitution or mere supervenience
>
> (*ibid.*, page in book needed)

They then go on to write:

> Neural activity seems capable of explaining only structure and function, but consciousness appears to have qualitative properties that are not purely structural and functional. Neurally-based explanations appear

> doomed to leave the 'what it's likeness' of conscious experience unexplained.
>
> The force of this worry is much disputed. … We will not take a position in this debate here, but will note that even if there are good reasons for thinking that certain aspects of consciousness are purely qualitative (and thus beyond the reach of neurally-based explanation), there may be other, interesting aspects of consciousness that are structural and functional.
>
> (*ibid.,* page in book needed)

In sum, Hohwy and Bayne recognize that there may be a fundamental problem for the NCC project but do not wish to let that get in the way of much interesting work which can be done without facing that problem. Hohwy's attitude to PEM relative to the problem posed by Putnam seems to be much the same. I shall be arguing that there need be no background worries of this sort for cognitive science because both problems can be resolved in one fell swoop.

The problem which 'what it's likeness' may pose for the NCC project can be captured in the idea of a 'zombie', which has prompted much discussion, as is well described in Kirk (2015). The idea can be put like this. I have no doubt that there is something it is like to be me, now. But it can seem conceivable and perhaps possible that there exists a distant region of the universe which is physically isomorphic to this region and in that other region I have an exact doppelganger which behaves exactly as I do but which lacks an 'inner life'. There is something it's like to be me but nothing it's like to be my doppelganger. The force of the zombie idea is, as Hohwy and Bayne write, 'much disputed', but if that force is real the

implication is clearly that the NCC project cannot succeed because there must be something more to being conscious than just having some neuro-functional structure.

Clearly there is a link between the concept of a zombie and Putnam's Twin Earth thought experiment. They both involve the idea of physically isomorphic doppelgangers. And both arguments share a ubiquitous assumption about the relation between the minds and the doppelgangers. It is that for each doppelganger there is a mind. It is not disputed that a zombie has a mind; its behaviour is exactly the same as its doppelganger's so mental contents can be attributed to both to explain those behaviours. The suggestion is that a zombie's mind is wholly unconscious; that there is nothing it's like to be a zombie.

For Putnam's doppelgangers on Earth and Twin Earth it is not their consciousness which is in question but rather an aspect of their mental content. Putnam suggests that the mental contents of the doppelgangers can be different if there is a difference in their environment and clearly the mental contents can only be different if the doppelgangers have numerically distinct minds. But although it is intuitively appealing to associate an individual mind with an individual doppelganger, it may not be necessary. It may be that physically isomorphic doppelgangers do not have numerically distinct minds; they share a single mind.

The idea was first explicitly expressed, so far as I know, by Gottfried Leibniz when he wrote. 'what is to prevent us from saying that these two persons who are at the same time in these two similar but inexpressibly distant spheres, are not one and the same person?' (1704, Bk.II, Ch.xxvii, 245). He is here considering two vastly separated isomorphic regions of the universe. An earlier hint is when Thomas Aquinas states that angels cannot be qualitatively identical and numerically distinct 'just as it would be impossible for there to be several whitenesses apart, or several humanities' (1274, Bk.I, Question 50, Article 4). More recently, Arnold Zuboff has considered the idea (1974, p. 374; 1991, pp. 41-2) and Nick Bostrom briefly discusses and dismisses it, of which more later (2006, pp. 186-88).

From a semantic internalist and neurocentric point of view the idea is that the brains of isomorphic doppelgangers are loci of instances of a particular neuro-functional type and it is that instanced type itself which is to be associated with the mind of a single subject, not its instances. I use the term 'associated' to leave open exactly how the relation between a mind and an instanced neuro-functional type is to be construed, whether 'by way of identity, constitution or mere supervenience', to use Hohwy's and Bayne's expression.

To be clear, what is being suggested is that if there were to exist multiple isomorphic doppelgangers, the minds present ought not to be counted along with doppelgangers. The number of minds present is equal to the number of particular neuro-functional types of structure present, which is 1. It follows that the possibility that a pair of doppelgangers should exist, both of which behave as if they had minds but only one of which is conscious does not arise. This 'unitary interpretation of mind' definitely excludes the possibility of zombies.

To be sure, if it could somehow be established in a compelling way that the existence of zombies really is relevantly possible then that would scupper the proposal. But as things stand there is no general agreement that zombies *are* possible. And if they are not possible the reason may be exactly that that is because atom-for-atom duplicates ought not be imagined as having numerically distinct minds. But can matched doppelgangers really be interpreted as sharing a single mind? I shall show that they can by applying the idea to Putnam's seminal thought experiment.

## 5. Set and Object

We have Oscar on Earth and Toscar on Twin Earth. And Putnam's argument has full force, as he acknowledges, when Oscar and Toscar are taken to be physically isomorphic doppelgangers (1975, p. 144). But if we are to adopt the alternative unitary interpretation of

mind then Oscar and Toscar are to be thought of not as people but as bodies which include two instances of a single mind. We have two instances of a particular complex neuro-functional type of structure 'in the head'. Let that one mind be the mind of Scar. Scar's mind has two instances, O and T, which are the relevant matched structures in the doppelgangers' heads. Scar's single mind spans two matched planets; there is an important difference between Putnam's Earth and Twin Earth but that can be set aside for the moment. And, practically speaking, we would not expect two planets to be matched unless their local cosmic environments were matched too. I shall say more about this later and in the meantime speak of 'parallel worlds'.

Imagine Scar faced with a green apple. He says 'I see a green apple before me'. Scar's single utterance is manifest as matched sonic emissions from the doppelgangers. If what he says is charitably interpreted as being true, what can the apple which Scar sees be? It cannot be the aggregate of the two matched 'parallel counterpart' apples since Scar tells us that his apple has a mass of a hundred grams and the aggregate of the matched apples has a mass which is twice that. Also, an apple cannot have apples as proper parts.

We are free to interpret Scar's apple as the *set* of the two parallel counterpart apples. Sets are normally thought of as abstract objects but concrete sets are feasible. This builds on a suggestion by Willard Van Orman Quine that an object such as an apple could be interpreted as a set which takes itself as sole element:

> none of the utility of class theory is impaired by counting an individual, its unit class, the class of that unit class, and so on, as one and the same thing.
>
> (1969, p. 31)

This suggestion is not generally regarded of being of any use to set theory, but there appears to be no telling objection to Quine's reconstrual of an individual as a set taking itself as sole element, which simply breaches the axiom of foundation, Forster (2006, p. 228). The apple Scar is faced with can be interpreted as the set of the two parallel counterpart apples, each of which is what logicians call a Quine atom, a self-membered singleton set. Interpreting individuals as Quine atoms makes an absolute distinction between mereology and set theory so, sad to say, David Lewis's 'Main Thesis' that 'the parts of a class are all and only its subclasses' must be rejected (1991, p. 7). Scar's apple's parts include its pips and peel, its subsets are the parallel counterpart apples since they are themselves sets.

Scar is faced with a doubleton apple, a set with two elements. And he reports that his apple is green. We can make sense of this if we suppose that he sees the apple as green because both its elements are green. That makes sense because if either one of the parallel worlds were zapped out of existence the only difference to Scar's mind would be that it would then have just one instance rather than two and Scar would be faced with the remaining green apple. Scar plucks his doubleton apple from a doubleton tree and munches it. His munching is manifest as matched mandibular motions.

Now comes the twist. There is a hidden difference between the parallel worlds. We are back in a time before chemists had discovered the molecular constitution of the colourless liquid raining regularly from above. Scar is oblivious to the fact that in one world it's raining $H_2O$ and in the other it's raining XYZ. Miraculously, this difference between the worlds has left them parallel in every other relevant way and we put the consequent anisomorphism between O and T down to philosophico-poetic license; the fact that O contains $H_2O$ and T contains XYZ is to be supposed not to have cognitive consequences, as with the difference between pentacles and septacles.

Scar quaffs a glass of the stuff. It is a doubleton glass of liquid constituted by doubleton molecules. But a doubleton molecule of the stuff Scar quaffs has one element which is $H_2O$ and the other which is XYZ. However, this difference is hidden from Scar and we are free to suppose that an object in his environment has a definite physical property if and only if all its elements have that property in common, like that green apple. In which case Scar's clear liquid has indefinite molecular constitution; it is neither $H_2O$ nor XYZ, even though it is limpid and thirst-quenching.

Then Scar takes up chemistry. The moment arrives when he at last performs the crucial test on a sample of what he calls water and, hey presto! Scar undergoes personal fission into Oscar who comes to believe that the sample is $H_2O$ and Toscar who comes to believe that it is XYZ. What has happened is this. As the doppelgangers move in concert whilst Scar performs the experiment, O and T, the two neuro-functional instances of Scar's mind, become subject to different stimuli leading to differing cognitive phenomena and so qualitatively different minds.

Personal fission has been a subject of much debate but it is not obviously unintelligible so does not, as things stand, provide a reason to reject this internalist interpretation of Putnam's setup. Ted Sider has provided a metaphysics of transtemporal identity which is well adapted to describing the relationship between identity and fission (2001, p. 201). Since Sider's stage theory has not so far met with a killing objection and remains part of contemporary metaphysicians' toolbox, it is legitimate to use it in describing what happens to Scar when he probes the molecular constitution of that infamous clear liquid.

According to stage theory, Scar has two 'future counterparts', Oscar and Toscar, to whom he bears the relation *will be*. Scar is neither Oscar nor Toscar but will be each of them, though he does not bear the relation *will be* to the pair, Oscar and Toscar. Scar will not become two people. And Scar is a 'past counterpart' of both Oscar and Toscar, two distinct

people each of whom bears the relation *was* to Scar. To be sure, stage theory has the odd implication that many people have worked at writing this very sentence, but that is arguably not intolerably odd since all those people are persons who I, now, was.

On this analysis Putnam's original Twin Earth challenge to semantic internalism evaporates. Prior to the molecular investigation Oscar and Toscar simply do not exist as people with numerically distinct minds. Back then there was just Scar for whom what he called 'water' had no determinate molecular constitution because its doubleton molecules had one element which was $H_2O$ and one which was XYZ. However, this liberation from Putnam's challenge to semantic internalism and from zombies arises out of an idea which many may find very hard to accept and which can have some bizarre consequences which I shall explore below.

## 6. Back to Reality

So, is this revisionary metaphysics of minds and their perceived environments really acceptable? Might there be good reasons to reject it? Putnam's Twin Earth is a fanciful creation and in that context a fanciful metaphysical response might not seem out of place but the unitary interpretation of mind can have some very counterintuitive consequences in the context of contemporary cosmology, which takes the existence of parallel worlds as physically possible, as the following paragraph explains.

In a nutshell, standard Big Bang cosmology has it that we inhabit a causally isolated region of space which is known as our observable universe. It is isolated because causal influence cannot propagate faster than the speed of light *in vacuo* and the time which has elapsed since the Big Bang is finite, so there can be distant objects which have had no causal influence on our local environment. Our observable universe has a finite volume and

according to quantum mechanics there is a finite number of definite physical states which that volume can possibly occupy. Since there is as yet no evidence that space and the number of galaxies it contains is finite, it is possible that there exist two or more regions of space which are physically type-identical to what we take to be our observable universe and have type-identical histories. These are spatially separated 'parallel universes', Tegmark (2007, p. 104). In what follows no simultaneity between parallel universes is presumed.

Suppose that there do in fact exist many instances of what we take to be our observed universe. On the conventional 'plural interpretation of mind' you are here in this universe and your doppelgangers are far off in the distance, in other regions of space causally isolated from our own[i]. On the alternative unitary interpretation your doppelgangers are not far off, they are all right here. The body which you perceive as yours is a set which takes the doppelgangers as elements. And any object in your observed environment, such as a green apple, is a set of parallel counterpart green apples. This can seem absurd; we start off by hypothesizing parallel universes which are far apart and then claim that they are all in the same place!

The absurdity is only apparent. The unification arises simply because the doppelgangers share a physical form, no other connection between them is being posited. Having left Scar behind in Putnam's imaginary setup let us now introduce a new character for the cosmologically respectable setup where there exist many isomorphic observable universes of any given type. Call her Hydra. Hydra's body is a set of dopplegangers: $\{H_1, H_2, \text{etc.}\}$. And Hydra can tell you where she is. Multiphonically, she says 'I'm 5,205 kilometres from the North Pole and 5.37 degrees east of the Greenwich meridian'. Hydra's North Pole is a set of parallel counterpart poles: $\{p_1, p_2, \text{etc.}\}$, and her Greenwich meridian is a set of parallel counterpart meridians: $\{m_1, m_2, \text{etc.}\}$. The distances $H_1$ to $p_1$, $H_2$ to $p_2$, etc., are all 5,205 kilometres and the angles $H_1$ to $m_1$, $H_2$ to $m_2$, etc., are all 5.37 degrees. According to the rule I introduced earlier, Hydra has the property of being 5,205 kilometres from her North Pole if

and only if each of the pairs of objects ($H_1$, $p_1$), ($H_2$, $p_2$), etc. have the property of being 5,205 kilometres apart. They do, so she does.

Likewise, Hydra can be taken to speak truly about her longitude. And a similar exercise with events and clocks allows her to be understood to speak truly about times, recalling the no simultaneity between parallel universes is being assumed. So the idea that there exists a multitude of parallel universes which are spatiotemporally separated is not at odds with the idea that your mind spans all of them in a perceived here and now because a perceived spatiotemporal region is to be interpreted as a set of parallel counterpart spatiotemporal regions.

But there is another reason why the unitary interpretation of mind can seem absurd. We attribute mental content to creatures in an attempt to explain their behaviours. If Hydra believes that an apple is crisp and juicy and so desires to munch it, that explains why she eagerly plucks it. But it seems obvious that we can explain the behaviour of each of the doppelgangers in exactly the same way. So the conventional plural interpretation of mind must be right after all. In each parallel universe the causal relation between the doppelganger and the apple is local, so the behaviour must be locally explained.

This brings us to the deep issue of the nature of causality. In everyday life we take causation to be a relation of natural necessitation. The red billiard ball moves thus and so because struck by a cue ball. David Hume is often taken to have shown that the very idea of natural necessitation is incoherent, though Galen Strawson has argued that Hume believed that there is indeed a 'secret connexion' between events, albeit as yet of an unfathomable nature (1989). Let us suppose to begin with that a real causal relation may exist between the impact of the cue ball and the rebound of the red ball. I shall consider afterwards the alternative 'regularity' view of causation which is often attributed to Hume.

Hydra's cue ball strikes the red. Her cue ball is a set of balls, as is the red. In each parallel world there is a cue ball and a red ball each of which is a Quine atom, a self-membered singleton set. If, somehow-we-know-not-how, there is a relation of natural necessitation between the striking and rebounding balls then it is a relation between sets. But then there is no reason to suppose that there is not a relation of natural necessitation between Hydra's striking and rebounding balls, which are also sets. So we can explain Hydra's behaviour by attributing mental contents to her just as easily as we can explain the behaviours of her doppelgangers if they are regarded as the bodies of distinct subjects. So the concept of causation as a relation of natural necessitation only presents a difficulty for the unitary interpretation of mind if that relation is taken just to exist between individuals which are non-sets. But then some principled objection is needed to Quine's suggestion that individuals can be interpreted as self-membered singletons.

Now consider causality as constant conjunction. On that view, to say that the striking of the cue ball causes the red ball to move thus and so is just to say that similar events have always gone like that. Since parallel universes are by hypothesis isomorphic and have isomorphic histories a constant conjunction of events in one universe is necessarily a constant conjunction of sets of parallel counterpart events. So, again, there is no reason to reject attributing mental contents to Hydra to explain her behaviours since here bodily movements are sets of parallel counterpart bodily movements. If, for whatever reason, the universes which Hydra's mind spans should cease to be parallel in ways which affect her mentality then, as we saw with Scar, she will undergo personal fission into distinct people with qualitatively distinct minds. Each of those persons' minds will span a subset of universes which have remained parallel. So each of those persons' histories will be a history of constant conjunctions between sets of parallel counterpart events. Whichever of the two views of causality is favoured, there is no reason to suppose that conventional plural interpretation of

mind is to be preferred to the alternative unitary interpretation, despite intuitions to the contrary.

## 7. Varieties of Divergence

Since we have returned to reality via contemporary cosmology we had better take quantum theory into account too, and that seems to create bizarre consequences for the unitary interpretation of mind if a multitude of parallel universes exists. Consider standard quantum mechanics which takes quantum processes to be stochastic. And suppose that Hydra's mind spans a very large or infinite number of parallel universes. She can make a measurement of the spin of a particle relative to some spatial axis, for instance, which quantum mechanics tells her has two possible definite outcomes on the 'pointer basis' indicated by a pointer which moves either left or right. Call these two possible outcomes L and R, with probabilities pL and pR. As Hydra makes the measurement a stochastic process takes place in each of the parallel universes which her mind spans. In some of the universes the result is L and in some it is R. So the original set of universes partitions into two subsets and Hydra fissions into two observers, $\text{Hydra}_L$ who sees L and $\text{Hydra}_R$ who sees R. If the sets of universes are infinite then the relative measures of the subsets on the original set are pL and pR, the 'probability measures'. If the sets of universes are large but finite the measures will be proportions and will approximate to pL and pR, given the Law of Large Numbers.

According to the conventional plural interpretation of mind, an observer in a single universe, prior to making the measurement, speaks truly when s/he says 'The probability that I will see the result R is pR' because there is an utterance in each universe referring to a stochastic process. But when Hydra speaks the sounds coming from the mouths of the many doppelgangers do not each voice an utterance. Hydra makes a single utterance which is

mulitphonic and her utterance refers to a process of partitioning, not a stochastic process. What Hydra refers to as objective quantum-mechanical probabilities are the relative measures of the subsets into which the pre-measurement set of universes partitions.

It is on the issue of probability that Bostrom makes what appears to be a telling point against the unitary interpretation of mind, which he calls 'Unification' (*op.cit.*). He imagines what he calls a Big World 'in which all possible human experiences are in fact made' (*ibid.*, p.187). Suppose that in that world the cosmic background temperature is 2.7°K, as *chez nous*. Observers whose evidence derives from a reliable sampling of the world will measure approximately 2.7°K. But there will be 'unlucky' observers who get an unreliable sampling so that they measure, say, 3.1°K. 'Yet', as Bostrom puts it, 'if Unification were true, then experiences of observing 3.1°K would be just as frequent as experiences of observing 2.7°K' (*ibid.*). He goes on to conclude, 'Thus Unification would undercut a natural account of why our experiential evidence enables us to learn about the world (even if the world is a Big World).' (*ibid.,* p. 188).

This can indeed seem intuitively compelling. However, Bostrom does not take into account the constitution of the unified subject's environment. What is at issue here is the probability that a subject making a measurement of some parameter has access to a representative sampling of the data. As we saw with Hydra, in an *infinite* world where there are different possible observations with associated probabilities, the post-observation subset measures for a fissioning subject just *are* the probabilities of the various observations in each 'branch' of the partitioning. For a non-infinite Big World corresponding branch measures will approximate to the probabilities, given the Law of Large Numbers. If $\text{Hydra}_{2.7}$ observing 2.7°K thinks she has reliable sampling, she's right, because she's on a branch with a high probability measure. But if $\text{Hydra}_{3.1}$ thinks she has reliable sampling, she's wrong, because she's on a low-probability branch. The unitary interpretation of mind is simply a change of

perspective. Bostrom's frequencies still indicate probabilities, given the Law of Large Numbers, but in a different way than they do on the conventional plural interpretation of mind which he defends.

There is another consequence of combining cosmological parallel universes, quantum mechanics and the unitary interpretation of mind which can appear troubling. Suppose that when Hydra initiates her quantum measurement she does not immediately observe the outcome because the pointer is facing away from her. However, Hydra's friend Kaliya who is sitting opposite can see the pointer. As the measurement takes place Kaliya fissions into $\text{Kaliya}_L$ and $\text{Kaliya}_R$ but there is no good reason to suppose that Hydra fissions. To be sure, the physical changes in the pointer and in Kaliya will have effects which propagate away from them in Hydra's direction and which will cause some changes in Hydra's body, but it is implausible that those very small changes would change Hydra's cognitive state, either conscious or unconscious.

The implication is that there can come to be properties of Hydra's own body which are indefinite for her since not all the doppelgangers which are elements of her body are physically isomorphic. Also, Kaliya's cognitive state will cease to be definite relative to Hydra since Kaliya has fissioned and Hydra has not. Whilst Hydra remains ignorant of the result after the measurement there is not a single person facing her but rather a human body which is a set of doppelgangers with subsets constituting the bodies of two people, $\text{Kaliya}_L$ and $\text{Kaliya}_R$. Only when those two people tell Hydra the results they see does Hydra fission into $\text{Hydra}_L$ and $\text{Hydra}_R$.

In a similar vein, consider Jack the Ripper (footnote to be added after blind review). His having had a wart on his nose, or not, may well be compatible with your exact cognitive state now, both conscious and unconscious. This is an extremely complex matter. Whether or not Jack had a wart on his nose will have myriad knock on effects which may influence

everything from the New York stock market in 1929 to your local weather as you read this. So it is not obviously true that your cognitive state is independent of Jack's wart, but it might be and we are unlikely to ever be sure that it is not. In which case you cannot be sure that your mind does not span universes which are only partly parallel, some of them containing in their past a doppelganger which was an element of Jack's body with a wart and some a doppelganger without. So you cannot be sure that there is a fact of the matter as to whether the person you refer to as Jack the Ripper had a wart on his nose or not. Even if Jack's body has been preserved and lies now in a Welsh bog its wartiness may be indefinite relative to you. In which case, if you were to dig it up, you would fission into persons finding and not finding a wart.

I should stress that strange consequences such as this only arise from the unitary interpretation of mind if there exist multiple parallel universes of the sort we observe which is plausible according to contemporary cosmology but which might not be the case.

## 8. Parting Lines

I have argued that the unitary interpretation of mind can dissolve the zombie problem and undermine Twin Earth arguments for semantic externalism. That leaves us free of two significant background worries whilst pursuing a physicalist/materialist cognitive science on the assumption that minds are wholly contained in heads, with the possible addition of sufficiently well integrated peripherals. There could be strange consequences if we inhabit a world of many parallel universes in which stochastic processes take place but I hope to have shown that this alternative view of the mind-body relation is worthy of consideration.

If the idea is indeed acceptable it makes the possibility of personal transport via teleportation and survival via mind uploading more plausible. For if the conventional view

that minds can be qualitatively identical and numerically distinct is accepted it would seem that teleportation and mind uploading can only ever lead to the creation of a person just like oneself but numerically distinct; a person who is no more you than is your doppelganger on a distant planet. Only if matched neuro-functional structures instance the mentality of one and the same person do teleportation and mind uploading appear possible as means of personal transport and survival.

## References


Aquinas, T. (1274), *Summa Theologica*. New York: Benziger Bros., 1948.

Bostrom, N. (2006), 'Quantity of experience: brain-duplication and degrees of consciousness', *Minds and Machines* 16:185-200.

Burge, T. (1979), 'Individualism and the mental', *Midwest Studies in Philosophy* 4:73–121.

Burge, T. (1982), 'Other bodies', in Woodfield, A. (ed.), *Thought and Object,* Clarendon Press, pp. 97-120.

Clark, A. (2016), 'Busting out: predictive brains, embodied minds, and the puzzle of the evidentiary veil', *Noûs,* doi: 10.1111/nous.12140.

Clark, A. & Chalmers, D. (1998), 'The extended mind', *Analysis* 58:7-19.

Curtis, B. L. (2015), 'On there being infinitely many thinkable thoughts: a reply to Porpora and a defence of Tegmark', *Philosophia* 43:35-42.

Davidson, D. (1987), 'Knowing one's own mind', *Proceedings and Addresses of the American Philosophical Association* 61:441–458.

Forster, T. (2006), 'Permutations and wellfoundedness: the true meaning of the bizarre arithmetic of Quine's NF', *The Journal of Symbolic Logic* 71:227-240.

Hohwy, J. (2013), *The Predictive Mind*, Oxford University Press.

Hohwy, J. (2014), 'The self-evidencing brain', *Noûs*, doi: 10.1111/nous.12062.

Hohwy, J. & Bayne, T. (2015), 'The neural correlates of consciousness: causes, confounds and constituents', in Miller, S. M. (ed.), *The Constitution of Phenomenal Consciousness*, John Benjamins.

Kirk, R. (2015), 'Zombies', *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/zombies/ .

Leibniz, G. W. (1704), *New Essays on Human Understanding*, Cambridge University Press, 1996.

Lewis, D. (1980), 'A subjectivist's guide to objective chance, in Jeffrey, R. C. (ed.), *Studies in Inductive Logic and Probability Vol. II*, University of California Press, pp. 263-293.

Lewis, D. (1991), *Parts of Classes*, Blackwell.

Porpora, D. (2013), 'How many thoughts are there? or why we likely have no *Tegmark duplicates* $10^{10^{\wedge}115}$ m. away', *Philosophical Studies* 163:133-149.

Putnam, H. (1975), 'The meaning of "meaning", *Minnesota Studies in the Philosophy of Science* 7:131-193.

Putnam, H. (1981), *Reason, Truth and History*, Cambridge University Press.

Quine, W. V. O. (1969), *Set Theory and its Logic*, Harvard University Press.

Russell, B. (1954), *Nightmares of Eminent Persons*, The Bodley Head.

Sider, T. (2001), *Four Dimensionalism*, Oxford University Press.

Strawson, G. (1989), *The Secret Connexion*, Oxford University Press.

Tegmark, M. (2007), 'The multiverse hierarchy', in Carr B. (ed.), *Universe or Multiverse?*, Cambridge University Press, pp. 99-125.

Zuboff, A. (1973), 'Nietzsche and eternal recurrence', in Soloman, R. C. (ed.), *Nietzsche: A Collection of Critical Essays*, Doubleday, pp. 342-357.

Zuboff, A. (1991), 'One self: The logic of experience', *Inquiry* 33:39–68.

[i] Douglas Porpora has argued that such doppelgangers, despite having a finite number of possible physical states, may entertain an infinite number of different thoughts. And his argument is not that that is so because they make indexical reference to an infinite number of numerically distinct environments (2013). A reply is to be found in Curtis (2015). In what follows I am assuming that Porpora's argument fails.