Could Consciousness Be an Illusion?

William Seager

September 26, 2016

1 Nonsense on Stilts?

The claim that consciousness could be an illusion seems preposterously, ridiculously false on its face. Is it not evident that if we know anything at all we know that there is consciousness, or that there are states of consciousness? If an argument is needed to bolster our confidence, it is not hard to provide one, which we might call the obvious argument:

(1) If consciousness is an illusion then it merely seems that it exists.

(2) But if anything *seems* to exist, that seeming is a state of consciousness.

(C) Therefore consciousness (states of consciousness) exists.

The obvious argument shows that there is evidently a basic difference between the case of consciousness and other aspects of experience that one might think are immediately revealed to us. Think, for example, of the question whether there is such a thing as temporal passage. There is little doubt that consciousness 'reveals' temporal passage insofar as the sense of flowing time is a core feature of our experience of time. Yet many philosophers and physicists have argued that this experience is illusory; in fact, the dominant view currently is that there is no such thing as flowing time. The argument presented above is not based on what is revealed in or to consciousness. It is not that the existence of consciousness is a feature of experience, which feature might or might not correspond to anything in reality. Rather, the idea is the undercutting one that if there is experience which has features *at all*, then consciousness must exist.

Nonetheless, confirming Descartes's claim that there is no idea so strange but that some philosopher has maintained it, efforts have been made to show that consciousness is nothing but an illusion. I want to consider some of the lines of thought that lead to this peculiar conclusion. Of course, they do not succeed, but they will be instructive both about our concept of consciousness and consciousness itself.

One of William James's apparently most provocative of views is foreshadowed in the title of his article 'Does "Consciousness" Exist?'. James does indeed begin with this blunt statement:

> I believe that 'consciousness,'... has evaporated to this estate of pure diaphaneity, is on the point of disappearing altogether. It is the name of a nonentity, and has no right to a place among first principles. Those who still cling to it are clinging

> to a mere echo, the faint rumor left behind by the disappearing 'soul' upon the air of philosophy (1904b, p. 477).

James is not really about to deny the existence of consciousness however. He immediately backpedals:

> To deny plumply that 'consciousness' exists seems so absurd on the face of it–for undeniably 'thoughts' do exist–that I fear some readers will follow me no farther. Let me then immediately explain that I mean only to deny that the word stands for an entity, but to insist most emphatically that it does stand for a function. There is, I mean, no aboriginal stuff or quality of being, contrasted with that of which material objects are made, out of which our thoughts of them are made... (p. 478).

What James is trying to get at is the core idea of his neutral monism according to which neither mind (as consciousness) nor matter are fundamental constituents of the world. Both of these familiar aspects of the world are conceptual constructs whose 'material' is the neutral stuff which James (misleadingly) calls 'pure experience' (see also James 1904a). Neutral monism strives to erase the subject/object distinction as ontologically significant. Experience does not testify to the existence of the conscious subject as a fundamental component of the world. This is, however, far from a denial that consciousness exists. That would be to deny that there is anything ever immediately present. What is arguably illusory for James is the existence of a substantial conscious self, for the self is simply another conceptual construct, a way, but not a mandatory way, to organize the flux of pure experience.

To find a more robust denial of consciousness we need to jump forward a good part of a century beyond James to the writings of Daniel Dennett (see e.g. 1991a, especially ch. 4, or 1978a). Dennett identifies a possible fallacy in the obvious argument in favor of consciousness. It can be claimed that the argument equivocates on the term 'seems'. This word has at least two distinct meanings. One is the 'seems' of experiential appearance, the other is the 'seems' of epistemic appraisal. It is of course extremely common to use the language of seemings where there is no relevant experience, as when one says something like 'it seems incredible that Trump is a presidential candidate'.

It is open, at least formally, to accuse the obvious argument of trading on this ambiguity. The 'seems' which supports the existence of consciousness is that of appearance, whereas, it may be argued, the sense in which it 'seems' that consciousness exists may be merely epistemic.

To make good this accusation, Dennett will need a way to explain epistemic seemings that does not in any way appeal to or invoke states of consciousness[1]. As we shall see below this is an impossible task, but the strategy does show that there is a way of denying the existence of consciousness without instantly falling into the absurdity revealed in the obvious argument.

There is a more subtle way to consign consciousness to the realm of the illusory. This way proceeds by claiming an analogy between, on the one hand, the illusion of a special kind or domain of knowledge that stems from *indexical* thought and language and, on the other hand, the supposedly illusory domain of phenomenal consciousness[2]. It is easy to fall into the belief that indexical knowledge must involve a special domain based on the seductive fact

that no amount of 'purely objective' non-indexical knowledge about the world will allow a subject to infer indexical knowledge.

Imagine you wake in a dark room with no memory of who, where or when you are. If you are restricted to asking questions limited to objective facts, it will be hard to locate yourself in any of those three dimensions. For example, one is allowed to ask 'What year did WW2 end' but not 'Is WW2 over'. Such thought experiments are impure because we are always forming indexical knowledge simply in virtue of being intelligently awake. Thus, no doubt your first instinct would be to ask something like 'how many people are in a dark room trying to discover their own identities'. If you are lucky the answer will be 'one' and you are off to the races. However, of course, this approach trades on your indexical knowledge that you, yourself, are in a dark room trying to discover your own identity, and the implicit indexical fact that the time in question is now. It can easily seem that there must be a domain of facts about one's own identity and about one's own space-time location which somehow go beyond the domain of objective fact. If this is in fact illusory, perhaps one could argue as well that the sense that phenomenal consciousness also constitutes a special domain that transcends that of the objective physical domain is similarly a kind of indexical illusion. Thus, just as I am (as I write this), just WS at 45.394359N, -64.233201W at 14:35 ADT on August 13, 2016, my current experiences are just, and no more than, certain objectively specifiable brain processes occurring at that time and place.

2 Motivations

Before considering these two latter approaches (James's is not a serious attempt to show that consciousness is an illusion) I want to consider the motivations that would lead a philosopher to embark on this strange project. It seems evident that there is a deep worry that consciousness cannot fit into a certain otherwise highly attractive view of the world, which we might call the scientific picture of the world[3]. According to the scientific picture, the world began in a relatively simple and purely physical state, constituted by a small number of simple physical entities (presumably various quantum fields) interacting according to the fundamental laws of physics. Modern cosmology has developed a remarkably complete account of the very early universe (from about a trillionth of a second after the Big Bang) whose analysis leads to a range of predictions, and explanations of known data, which have been verified (e.g. the ratios of light elements, the universal microwave background radiation, etc.). The general physicalist project is to give an account, as complete as possible, of how the features of the world which we now observe emerged from the initial conditions of the Big Bang. The metaphysics of physicalism holds minimally that everything is physical and emerged as determined by these initial conditions operating under purely physical laws which, it is held, modern physics has more or less completely cataloged. Of course, we know that physics is incomplete at present and struggles to incorporate all the known fundamental features, notably gravity, into a single theoretical account.

But a key idea bolstering the scientific picture is that the unknown physics which will, sooner or later, reveal to us the overarching physical theory of everything (everything fundamental that is) makes no difference to how things evolved starting about a picosecond after the Big Bang. This viewpoint is forthrightly expressed by the physicist Sean Carroll:

> The Laws Underlying The Physics of Everyday Life Are Completely Understood... All we need to account for everything we see in our everyday lives are a handful of particles—electrons, protons, and neutrons—interacting via a few forces—the nuclear forces, gravity, and electromagnetism—subject to the basic rules of QM and GR (2010).

Carroll is of course highly cognizant of all the problems we have integrating quantum mechanics with general relativity. His point is that these problems infest a region of physical phenomena which does not impinge upon 'everyday life'. In the low energy regime in which we live and have our being, the esoteric physics needed to explain those domains where gravity could affect microphysical events is irrelevant[4]. At the same time, all natural domains can be fully resolved into fundamental physical constituents. The physicalist holds that the properties and arrangement (initial conditions, boundary conditions) of the basic physical entities logically determine everything else.

Now, the standard understanding of the fundamental physical entities making up our world is that they are entirely devoid of consciousness. There naturally arises then, as part of the general physicalist project, the particular task of of explaining how it is that a world made of entirely non-conscious physical components generates consciousness. The project is at present a patchwork of more or less complete accounts of how the denizens of the everyday world arise, but in all areas but one the broad outlines of how the account should go are fairly clear and highly credible.

This catalog of success could be extended indefinitely, up to and including the large scale structure of the universe and the physical nature of life itself. It is important to be clear about the scope of the physicalist project. It is not required that all domains be *reduced* to that of fundamental physics, in any traditional sense of reduction. Here we can deploy the distinction between ontological and theoretical reduction. The latter is a system of relations between theories which permit derivability of reduced theories from the reducing theory or theories and thus the in principle elimination or supercession of the non-fundamental theories. The theoretical reductionist dream, which now seems unrealizable, was to reduce all theories to fundamental physics[5]. The former view is the conjoining of the claim that the fundamental physical state of the world logically determines everything with the claim that there are no brute logical necessities. Brute logical necessities are necessities that hold with maximal modal force but which are neither intrinsically intelligible nor follow from more basic necessities[6].

To see the significance of brute necessities, consider a view such as that held by the so-called British Emergentists of the early 20th century[7]. According to this view, when fundamental physical entities arrange themselves in particular configurations or relational structures there arises (emerges), as a matter of natural law, a new feature of the system. This feature could not, even in principle, be accounted for in terms of the properties and arrangement of the fundamental physical constituents of this special relational structure. This new feature is causally efficacious and the physical system will, post emergence, act differently than would be predicted by considering fundamental physics alone. However, it is merely a law of nature that links the physical relational structure to its emergent feature. While it is *nomologically impossible* for the emergent feature to fail to appear when the relational structure arises, it is possible in the broad sense that the emergent not appear.

There are, that is, nomological brute necessities, that is, laws of nature that are neither intrinsically intelligible nor logically determined by other natural laws. To take one example, in our world a host of laws of nature involving electromagnetic interactions depend upon the fine structure constant, a dimensionless number which 'calibrates' the strength of these interactions. The value of the constant is very close to $^1/_{137}$. Nothing seems to necessitate this value. Its value has engendered a great deal of speculation. If it was much different, life as we know it could not have evolved in the universe. Richard Feynman waxed somewhat poetic about it: 'We know what kind of a dance to do experimentally to measure this number very accurately, but we don't know what kind of dance to do on the computer to make this number come out, without putting it in secretly' (1985, p. 129). Remarkably, there is now some evidence that the constant's value changes over cosmological scales (see Webb *et al.* 2011). Webb et. al. claim that their data suggest our observable 'universe' is a relatively small region of a much larger or infinite universe in which physical 'constants' may vary across time or space. This line of thought strongly supports the claim that it is at least possible for these perhaps universal constants to be different than they are[8].

The lack of any intelligible route from other fundamental physical features to the specific observed value of the fine structure constant underpins our sense of modal variation. In the realm of absolute necessity, lack of intrinsic intelligibility, such as enjoyed by the absolute necessity of, say, '$2 + 2 = 4$', or no logical connection to such basic necessities (as in the absolute necessity of 'given the laws of nature, water boils at 100°C at standard pressure') points to modal variation. It does not point to some weird domain of brute and inexplicable absolute necessities.

If there are no brute absolute necessities then physicalism requires that there be some intelligible relation between the basic physical constitution of the world and everything else that we ought to, or must, count in our ontology. And while virtually all aspects of the world do indeed present themselves as intelligibly fitting into the physicalist project, consciousness does not[9].

The classic anti-physicalist arguments all exploit this absence of any intelligible connection between, on the one hand, the postulated entities of fundamental physics, the interactions they participate in and the resultant structures they can form and, on the other hand, the fact that consciousness, or a 'subjective' aspect of nature, exists. I will assume here that these arguments point to a real problem for the physicalist project[10]. If we take these arguments seriously, as there venerability suggests we should, then there are two basic responses the physicalist can make. One is to beg for more time, more scientific development, in the hope that eventually the mystery of consciousness will evaporate in something like the way the 'mystery' of life (with its attendant anti-physicalist theory of vitalism) disappeared as the connection between living things and the physical processes of biochemistry, the structure of DNA, etc. became apparent. The difficulty with this quietist response is that there are no signs that or how consciousness will be brought into the physicalist fold.

The second, somewhat desperate, response is to deny the existence of the troublesome domain. The problem of consciousness now seems so formidable that this second approach has been seriously contemplated. It is time to look more closely at the prospect of eliminating consciousness.

3 Dennett's Intentional Irrealism

Daniel Dennett's views on consciousness and cognition, primarily developed in *Consciousness Explained*, (1991a), but see also (2005, 2001a, 2001b), are complex and multi-faceted. On the side of cognition, Dennett has developed a pandemonium[11] model, in which a vast set of contentful brain processes (the 'demons') vie for control of behavior, including the all important speech behavior. The demons are selfish but political[12], so are willing to enter into temporary and uneasy coalitions with one another in the furtherance of their 'aims'. It is said that every age models the mind on current impressive technology (the mechanical clock, the steam engine, the telephone switchboard, the computer). Dennett presents us with a bureaucratic model of the mind, as may sadly befit the 21st century academic and scientific enterprise.

This model which sees the brain's implementation of cognitive functions as highly disunified and internally conflicted is highly interesting and not without some plausibility. But our question is: how or where does consciousness fit into this picture? On the face of it, the Dennett's model does not seem to constrain us to any particular view of consciousness.

Dennett is a presumptive physicalist, though he has not spent any appreciable time arguing in favor of it, who would see himself as a participant in the physicalist project. But there is nothing in the pandemonium model itself with reveals the ultimate purely physical nature of consciousness or how consciousness is strictly determined by the physical. And Dennett can see the tensions that exist between physicalism and the strange fact that nature appears to possess both a subjective as well as an objective aspect. His solution is to claim this is mere appearance:

> There seems to be phenomenology... But it does not follow from this undeniable, universally attested fact that there really is phenomenology (Dennett 1991a, p. 366).

Now, here Dennett invites the a reply based upon the obvious argument. And as we have seen, there is a response available which is that the argument exploits an ambiguity in the term 'appears' or 'seems' which has both a sense which involves subjective experience and a sense which is merely epistemic.

If it is possible to interpret 'there seems to be phenomenology' in the epistemic sense then consciousness can be demoted to a non-existent intentional object and the acknowledged difficulty in integrating consciousness into the scientific-physicalist picture of the world will evaporate.

To bolster his interpretation, Dennett develops an account of a kind of philosophical anthropology, called 'heterphenomenology', which permits a 'neutral' assessment of claims about consciousness. It works like this:

> [heterphenomenology] involves extracting and purifying texts from (apparently) speaking subjects, and using those texts to generate a theorist's fiction, the subject's heterophenomenological world. This fictional world is populated with all the images, events, sounds, smells, hunches, presentiments, and feelings that the subject (apparently) sincerely believes to exist in his or her (or its) stream of consciousness (Dennett 1991a, p. 98).

Dennett's conceit is that we should regard claims, even our own claims, about the existence of subjective consciousness as akin to the (sincere) stories told by a distant tribe which we (conscious beings?) are visiting to investigate their peculiar mythologies (this strange attitude is even more evident in Dennett 1978a).

It is worth pausing to consider how flummoxed one would have to be by the problem of consciousness to venture onto Dennett's path. Suppose we visit a distant people in a strange country and, after we learn their language, they tell of strange, terrible and deadly creatures that stand ten feet tall and weight 800 pounds. These creatures sometimes marry women and so must not be eaten, but must be respected by leaving meat from successful hunts, and so on. We might think that this is a curious myth, until we come upon a grizzly bear towering above us. Of course, we might still doubt the marriage tale. In the gradation from myth to legend to truth, the first two categories should not be conflated.

In the case of consciousness, why would we not believe our 'apparently' speaking subjects? We too recognize states of consciousness within ourselves, the same as they seem to be talking of. We might not believe everything we hear about it (e.g. consciousness resides in an immaterial spirit), but incomplete or partial knowledge is not the same as wholesale error. One can only conjecture that there is a great fear that consciousness, once admitted to any degree, will threaten to collapse the physicalist project.

In any event, to pursue the heterophenomenological method requires assigning meaningful content to the utterances of our 'informants'. This assignment cannot, or course, itself depend upon any appeals to states of consciousness. Dennett famously holds an interpretationist theory of content according to which meanings are determined via the 'intentional stance' (see 1971). The intentional stance is basically the deployment of belief-desire psychology in the effort to explain and predict behavior. To apply it, one makes assumptions about what sort of things one's target is likely to believe and desire given the environment of interpretation, and then, assuming some basic level of rationality, predict the target will do something which would, should the target's beliefs be true, lead to satisfaction of some of the target's desires. Throughout the realm of living things and complex artifacts, the intentional stance is remarkably powerful[13]. By and large, it is the only practical way we have of predicting what other human beings and most animals will do.

In the case of heterophenomenology, our interpretation is to aid in the prediction and explanation of what our targets say about consciousness. If talk of consciousness, phenomenal character and subjective phenomenology can be used in an intentional stance interpretation of a target, as we know it can on the basis of our own philosophical and everyday talk, then this is what we should regard the target as thinking and talking about. But our interpretive success does not entail anything about the reality of the 'things' our targets are talking about.

But, as noted, when we ourselves more or less agree with and have experience in concordance with our targets' discourse there seems to be little reason to dispute its general accuracy. Furthermore, there are fundamental problems with Dennett's interpretationism when it is over-extended enough to underwrite the heterophenomenological project.

In the first place, for heterophenomenology to do its undermining work we have to be able to apply it to ourselves. Yet it is very doubtful that our knowledge of our own conscious states involves anything like a self-applied intentional stance. While it is arguable that our knowledge of our own character and personality traits is based upon self-interpretation, a viewpoint that goes back at least to Gilbert Ryle (1949[14]), there is no plausibility to the

idea that our access to the *existence* of states of consciousness is similarly self-interpretation dependent. The conscious states which I enjoy are the necessary starting point for my interpretive endeavors, or self or others.

Secondly, if taken as a general account of mind in the world, Dennett's account seems to be viciously circular, since the concepts involved in the intentional stance are obviously mentalistic in nature. Here one must adopt a kind of transcendent counterfactual position by which *interpretability* by the intentional stance provides a position which does not presuppose the existence of thinking subjects. According to this line of thought, which Dennett develops in 'Real Patterns' (1991b), interpretable patterns of behavior are the metaphysical bedrock grounding the existence of mental states. These patterns are objectively present, ready to be interpreted via the intentional stance if the occasion, and the interpreters, should arise. But such patterns retain a conceptual dependence upon our notions of the intentional mental states and hence do not support the complete naturalization envisaged by the physicalist project.

Dennett may not be particularly troubled by this aspect of his view. He has shown little interest in the metaphysics of physicalism, describing himself as '. . . a reluctant metaphysician . . . [optimistic] about the innocence of the standard inventory of what we might call the ontology of everyday life and engineering' (2002, pp. 222-3). However, such a 'metaphysical quietism' seems at odds with the heterophenomenological rejection of consciousness, surely a commonplace part of everyday life (and engineering for that matter).

In fact, Dennett's pattern based metaphysics is quite radical. The being of patterns has a kind of conceptual relativity, dependent upon ways of experiencing distinctive of those who recognize the patterns. Dennett writes: '[w]ere there dinosaurs before H. Sapiens came along?. . . Of course there were, but don't make the mistake of thinking that this acknowledges a fact that is independent of H. Sapiens' (2002, p. 226)[15].

Given the uneasy relationship Dennett's views have to the physicalist project it is unclear why the rejection of consciousness via the heterophenomenological strategy needs to be endorsed. Dennett has explored a number of perplexing features of conscious experience which he may regard as indicating that the concept of consciousness is incoherent (see 1988, 1992 in addition to 1991a), but even if we accept the arguments against qualia and determinate temporality of experience we are only left with a modified or improved conception of consciousness, not anything like its wholesale rejection[16].

Lastly, I think a rejection of consciousness via the heterophenomenological method leaves us in an impossible intellectual position. If consciousness does not exist then what grounds the indisputable validity of the inference that *something* exists? It is simply obvious that we have contact with something and this contact either is or is mediated by consciousness. If we take the heterophenomenological method to its limit, why not declare that existence itself is an illusion. Sure, our subjects talk about things which exist, and sometimes about things that don't exist. To echo Dennett: sure, there seems to be existence. . . But it does not follow from this undeniable, universally attested fact that there really is existence. This is an incoherent extreme, but why does it fail? Because we are conscious beings appreciative, in this case, of a basic consequence of our own consciousness.

4 Indexical Irrealism

The claim that consciousness is an unusual kind of 'indexical illusion' can best be approached by returning to a classic anti-physicalist argument: Frank Jackson's 'knowledge argument' (1982). The argument is extremely well known, as is its 'spiritual progenitor' which is of course Thomas Nagel's (1974) famous lament that the nature of the subjective experience of creatures sufficiently different than human beings–and one does not have to go very far away from humanity–is more or less unknowable despite the openness to investigation of such creatures' physical state. Recall what Nagel says about trying to put oneself into the space of bat consciousness:

> ...it will not help to try to imagine that one has webbing on one's arms, which enables one to fly around at dusk and dawn catching insects in one's mouth; that one has very poor vision, and perceives the surrounding world by a system of reflected high-frequency sound signals; and that one spends the day hanging upside down by one's feet in an attic. In so far as I can imagine this (which is not very far), it tells me only what it would be like for me to behave as a bat behaves. But that is not the question. I want to know what it is like for a bat to be a bat (1974, p. 439).

Nagel is appealing here to a very powerful intuition, of which Jackson's knowledge argument can be regarded as a codification. In fact, the intuition is so pervasive and immediately moving that the knowledge argument as shown up, explicitly, in the popular film *Ex Machina* (of course, Hollywood does not get the argument and its point exactly right, and neglects to credit philosophy for the argument, but that doesn't matter here). Unsurprisingly, the core intuition is not new. John Locke expressed it thus:

> ...if a child were kept in a place where he never saw any other but black and white till he were a man, he would have no more ideas of scarlet or green, than he that from his childhood never tasted an oyster or a pineapple has of those particular relishes (1690/1975, Bk 2, Ch 1, §6).

Locke does not develop this thought into any philosophical argument, let alone an anti-physicalist argument. He seems to take it as utterly obvious that the qualities of experience cannot be known except via the having of the experience. Frank Jackson did present an argument however which proceeds via thought experiment. We imagine, and take it to be possible, that a future neuroscientist–named Mary–could come to know all the physical information relevant to color vision (we need not go so far as to make Mary physically omniscient unless and to the extent that understanding consciousness involves knowing absolutely everything about the physical universe). The argument then proceeds:

1. Mary has all the (relevant) physical information.
2. Mary nonetheless does not know what it is like to experience color.
3. Therefore, Mary does not have all the information there is to know about color experience.

4. Therefore, there is extra-physical information (and physicalism is false)[17].

Jackson's argument more or less implicitly assumes that Mary could not figure out or deduce what color experience is like from her abundant stock of physical information. How do or could we know that she has this limitation? No physicalist would hold that information about what experiences are like is a part of physical science as such. The physicalist holds that physical information entails all information.

Now, there is a trivial impediment to the project of Mary deducing information about color experience from her purely physical information. New concepts appear in the putative conclusion, but they cannot appear there unless they already figure somewhere in Mary's premises. And they cannot, non-trivially, appear in the premises because she has, by hypothesis, no color experiential information at her disposal. Clearly, the argument aims at a more substantial target.

Compare the project, already successfully in hand, of physically understanding water. On the one hand we have all the basic physical information that is relevant: information about oxygen, hydrogen and the quantum mechanics underlying chemical bonding. What we need to show is that H_2O instantiates the concept of water. The trivial impediment would arise if we did not possess the concept of water in the first place. The appropriate test of physical naturalization is this: *given* possession of the requisite physical information *and* the concept of water, can we show that this concept will be instantiated by H_20? This, it seems, we can do.

To test fully Jackson's scenario we ought to give Mary all the physical information and also give Mary the target concept. But then Mary won't be lacking any information at all, contrary to the letter of the argument. But a modified scenario abides, and in fact better abides, by the spirit of the argument. For what is important is the *intelligibility* of the link between the physical and the experiential which is the hallmark of the physicalist project.

So let us somewhat revise the thought experiment[18]. Mary has never seen color but has a normal brain and visual system in addition to her stupendously compendious store of physical information as well as a spectacular level of intelligence, mathematical acumen, vast computational resources, etc. We produce in her, perhaps via direct neural stimulation, two color experiences, one of a shade called R1, one of R2. She is then able to imagine them distinctly, able to remember them and would certainly recognize them correctly as either R1 or R2 if presented with them. She knows, in short, what it is like to experience them. She has, if you like, experiential or phenomenal concepts of R1 and R2. Let us say that Mary names the *experiential* quality of seeing R1 and R2, Q1 and Q2 respectively.

Now, Mary is a vastly competent and knowledgeable scientist. We imagine she has access to almost all physical information. She knows the laws of nature and the properties of matter. She fully understands how the brain, in particular, works. But in this version of the thought experiment she does not have total physical knowledge. She has no knowledge of the state of her *own* brain. We can however suppose that she has an exact twin. Mary does have complete knowledge of her twin's brain and its interactions with the environment.

Now, the question is can Mary figure out whether she will experience Q1 or Q2 when she is shown a sample of R1 (in standard lighting etc.). She knows or can figure out what state her brain will go in. But it does not seem that this will reveal which shade of red will be evinced by the sample, even though she knows exactly how the sample reflects light, etc. She can,

for example, figure out how many discriminable shades of color the brain can discriminate under optimal conditions (current estimates suggest between one and ten million different colors). That, given the rest of her knowledge, will let Mary figure out how many perceptible colors lie between Q1 and Q2 but won't reveal which quality will appear when she sees the R1 sample. Now, maybe R2 looks 'darker' and that is a function of the different reflectance dispositions of R1 and R2 but Mary won't know that the 'darker' *quality* goes, generally, with less reflectance. Although she possesses both sides of the equation, so to speak, she cannot show the connection which necessitates the shade of red she will experience given the brain state she knows she will go into.

Here, however, is a glitch. Given her physical knowledge, Mary can deduce what brain state she will go into when she sees a tomato. Can she not deduce from that the behavior that will result? In particular, can she not deduce that she will utter the words: 'now I know what it is like to see R1 (under that label) and it is Q1'? (Or, she could do this using the brain state of her twin, assuming that the twin has gone through the very same procedure.) And from that deduction she will have figured out which of Q1 or Q2 will be experienced when she sees the tomato.

But this deduction depends on some inductively attained knowledge of the link between seeing the tomato and Q1. It is simply a report of the linkage, the same as Mary herself will attain when she sees the tomato. It is not a 'pure' deduction from the physical state. That is, without the report Mary could not assign Q1 as the experiential quality which is what it is like to see R1, and yet the report does not include any information about this quality as such. It is nothing more than a second hand report of the same correlation Mary could (or will) establish between her own brain state and the experiential quality Q1.

It is interesting that this version of the thought experiment has to forbid Mary from having knowledge of her own brain states, even as it allows for her to have knowledge of her twin's qualitatively identical brain states. This imposed impediment irresistibly leads to a potential loophole in the anti-physicalist argument which is not philosophically unfamiliar. There is another kind of knowledge that shares with phenomenal knowledge the same peculiar feature that it is, apparently, not deducible from an in-some-sense complete body of knowledge: indexical knowledge.

John Perry (1979) noted that it seems that no matter how much 'objective' knowledge one has, this will not serve to place oneself in the world at the present time or current location, or even for one to self-identify. In another paper, Perry provides a fanciful example of the difficulty:

> An amnesiac, Rudolph Lingens, is lost in the Stanford library. He reads a number of things in the library, including a biography of himself, and a detailed account of the library in which he is lost. He... won't know who he is, and where he is, no matter how much knowledge he piles up, until that moment when he is ready to say, '*This* place is aisle five, floor six, of Main Library, Stanford. *I* am Rudolph Lingens' (1977, p. 492).

As with the Mary intuition, the recognition of the oddity of indexical knowledge is not novel. Immanuel Kant expresses the problem as one of personal orientation, using an example of literal self-location:

> . . . even with all the objective data of the sky, I orient myself geographically only through a subjective ground of differentiation. . . I also need the feeling of a difference in my own subject, namely, the difference between my right and left hands. . . I can. . . orient myself through the mere feeling of a difference between my two sides, the right and left. That is just what happens if I am to walk and take the correct turns on streets otherwise familiar to me when I cannot right now distinguish any of the houses (1786/1998, pp. 4-5).

Somewhat later, William James also recognized the issue and, curiously, linked it to the ineffability of phenomenal knowledge:

> If we take a cube and label one side top, another bottom, a third front, and a fourth back, there remains no form of words by which we can describe to another person which of the remaining sides is right and which is left. We can only point and say here is right and there is left, just as we should say this is red and that blue (1887, p. 14).

What is the relevance of the peculiar features of indexical knowledge to the knowledge argument? It stems from the sense one almost irresistibly has that there is some substantial gain in our state of knowledge when we, as Kant puts it, orient ourselves or in general 'locate' ourselves relative to the present time, current location and identity. We have all had the experience of losing our 'sense of location' (for me, this frequently happens when I emerge from a subway station on to an unfamiliar street corner) and this can seem to be, or largely involve, an epistemic lack. When we succeed in orientation, it certainly feels as if we have gained knowledge.

In general, new knowledge opens up new domains of facts. If indexical knowledge is ordinary knowledge there ought to be a corresponding domain of 'indexical facts'. But it seems quite clear that there is no such domain. One way to argue for this is to appeal to possibilities that arise from variation across independent domains of facts. If it is impossible to deduce indexical knowledge from non-indexical, then there should be modal variation of the former relative to the latter. This is of course completely analogous to the metaphysical situation suggested by the knowledge argument, where it seems that qualitatively different states of consciousness are compatible with identical physical states (or worlds).

But there is little or no reason to accept the existence of a domain of indexical facts. It makes no sense to imagine, for example, two possible worlds identical in every respect except for the temporal location of *now*. Similarly senseless are worlds where the location of *here* varies without any other variation[19]. Or that one's identity shifts across worlds that are identical in all other respects. At the very least, it appears that even if indexical facts are not reducible to the non-indexical, they are logically supervenient upon them. There are absolutely no possible worlds that are identical across every non-indexical feature but which differ about some indexical fact. This looks to be a promising way to think about the knowledge argument. Maybe phenomenal consciousness is logically supervenient upon the physical but it merely seems otherwise because of a kind of 'indexical illusion'.

Mary's situation is at least superficially similar to that of Lingens. We could write:

> A super neuroscientist, Mary, is wondering whether she will experience Q1 or Q2. She knows exactly the physical properties of the light she will be exposed to, she

> knows all there is to know about how light interacts with her visual system and she knows exactly what state her brain will be in at the time of the experience. She won't know what she is experiencing, no matter how much knowledge she piles up, until that moment when she is ready to say, '*This* is R1'.

Her indexical expression 'this' refers to a complex but perfectly ordinary physical state of her brain. Her rush of 'new' knowledge is not an indication of a new domain of fact but rather her successful orienting herself in the 'space of neurological states', as her rush of recognition of her location when she emerges from the subway is not awareness of a new fact but simply her successful orientation within physical space.

Now, normally we can all figure out who we are, when we are and where we are on the basis of preexisting knowledge and preexisting orientation. About consciousness, Mary somehow cannot orient so there still seems to be something special about her situation. But in fact there are cases of failure of ordinary orientation in well documented, if rare, instances of a variety of disorientation syndromes. Among these are 'topographic prosopagnosia' (also known as 'landmark agnosia') which is described as an 'inability to use prominent, salient environmental features for the purposes of orientation' and the more interesting 'egocentric disorientation' in which 'patients... recognize landmarks but cannot find their way because they do not know how to orient the body with respect to these landmarks (see Aguirre and D'Esposito 1999, p. 1617). There are remarkable examples of people suffering from these syndromes being unable to find there way home if they stray from memorized patterns (either walking or driving). Some sufferers have to put a super noticeable 'landmark' on their home (one woman used a giant lobster lawn ornament). Some cases of these syndromes are the result of an insult to the brain, but others appear to be idiopathic, perhaps congenital. An examination of one such case reveals the baffling inability to follow a map even when it is held in hand for continuous use as the subject attempts to get to a specified destination (see Bianchini *et al.* 2010).

It is tempting to diagnose the condition of those who fall prey to Jackson's knowledge argument as a kind of intellectual disorientation syndrome, which we might fancifully call 'developmental egocentric phenomenal disorientation' (the inability to orient one's current phenomenal state of consciousness with its neural realization). And perhaps there is a therapy for this condition. Perhaps if one became used to regarding one's states of consciousness from a neuroscientific perspective one would lose the sense of disorientation we are postulating as the explanation for the appeal of the knowledge argument. There is even a philosophical therapist available. Paul Churchland urges us to adopt neuroscientific language for self-description. Churchland holds of Mary that

> one test of her ability in this regard would be to give her a stimulus that would (finally) produce in her the relevant state (viz., a spiking frequency of 90 hz in the gamma network: a 'sensation-of- red' to us), and see whether she can identify it correctly on introspective grounds alone, as 'a spiking frequency of 90 hz: the kind a tomato would cause' (1985, p. 26).

I agree with Churchland that Mary could do this. But she would not thereby know whether she will experience Q1 rather than Q2. Obviously, there is no particular problem (beyond an insane cumbersomeness) with replacing our current language with a bunch of neuroscientific

words. The indexical approach suggests that Mary will still have to 'orient' herself with respect to her brain state and her current state of consciousness. Lingens won't be helped to locate himself by talking in GPS coordinates. Still, it would be cool to know what your brain was doing just by introspection and presumably, once neuroscience identifies the correlates of consciousness, we'll all be able to re-describe the qualities of our experiences in correct neural terms. This will not, however, solve the orientation problem, nor would a defender of the indexical knowledge approach think it would.

The core objection against the indexical knowledge approach is that the analogy is too shallow. There is a superficial similarity between the cases of Lingens and Mary, but it is no more than superficial. In standard topographic disorientation disorders both the objective and subjective representations are manifestly of the same domain of physical objects arrayed in space. It is this sameness that permits that 'rush of knowledge' when orientation is achieved. In the Mary thought experiment there is just as obviously an evident domain of proprietary qualities of experience. And as we have already seen above, it is not easy to simply deny that this domain exists or is a cognitive illusion. The distinctness of these domains prevents orientation between them, but of course does not prevent empirical coordination. Ordinary cases of indexical knowledge require coordination between different perspectives on the same domain. For example, in the case of ascending from the subway to find that one is disoriented, one gains the indexical knowledge of where one is by coordinating current indexical knowledge of the local topography with an objective representation of the same topography that one already possesses. The same topography exists in both representations. Coordination between neural and phenomenal states fails to engender orientation because of the distinctness of the relevant domains.

The knowledge argument depends upon the existence and accessibility of this proprietary phenomenal domain. I have argued above that this domain cannot be relegated to the dustbin of cognitive illusion. The indexical knowledge approach fails because it misleadingly assimilates coordination of representations of the same domain with coordination across domains. There are coordination problems lurking in the Mary thought experiment, but they are quite unlike those that underpin indexical knowledge.

Thus the two ways of arguing that consciousness is illusory do not seem at all promising. Consciousness remains in existence and remains problematic for physicalism.

Notes

[1]In what follows, I am going to assume (which I think is probably false) that epistemic seemings do not have their own phenomenal character. If they do, the equivocation charge against the obvious argument will founder on a second-order application of the argument to epistemic states themselves. The thesis that 'purely cognitive' states have a distinctive phenomenology is controversial so Dennett's attack on the obvious argument does not immediately fail (for defense of cognitive phenomenology see Pitt 2004).

[2]The approach as I will present it is inspired by work of John Perry (e.g. 1979; 2001) and Robert Stalnaker (2008), though I won't analyze their specific lines of argument.

[3]It might be more fair to call it the scientific-physicalist picture, but that is rather cumbersome, and the vast majority of those which embrace science as the basic authority about ontological matters also embrace physicalism. But it is worth remembering that one can value science without endorsing physicalism.

[4]The mathematical basis for his confidence is effective field theory, a formal methodology deploying so-called renormalization group procedures for developing and/or justifying theories which are empirically correct at relatively low energy–the energies our experiments can probe. As we build devices that can generate higher energy phenomena, or find ways to observe natural high energy events, our theories may begin to fail and 'new physics' may begin to appear. But this failure will not impugn the empirical adequacy of our old theories within their appropriate energy domain (see Cao and Schweber 1993 and Castellani 2002).

[5]See Suppe (1977) for a detailed account of the historical fortunes of classical reductionism

[6]Brute necessities are also known as 'strong necessities' (see e.g. Chalmers, 2009, who argues that there are none; further interesting considerations against brute necessities can be found in Fine 2008).

[7]The British Emergentists included such thinkers as John Stuart Mill, Lloyd Morgan and C. D. Broad among others. For a thorough review of their brand of emergentism, see McLaughlin (1992).

[8]A complication arises here if one holds that the causal powers of properties are definitive of those properties; that any two properties that differ in causal powers are different properties (for defense of such a view see Shoemaker 1980, Heil 2003). Proponents of such a viewpoint can hold that any physical duplicate of our world will be identical to our world tout court, even if there are non-physical entities. Emergentist powers will not vary across worlds that have identical physical properties. It seems to me an implausible view insofar as it seems fairly evident that mass, for example, could gravitate with force inversely proportional to distance raised to power 2.0000003 and still be mass. In any case, the core difficulty for physicalism will remain because it seems that there could be very similar physical properties that lack the power to produce the non-physical entities in question. The physicalist takes it that we are in a world with just *those* physical properties. Furthermore, the physicalist holds that everything is determined by the fundamental physical features of the world and it is very unorthodox to hold that the fundamental features can generate non-physical entities or possess intrinsically non-physical properties (a panpsychist, by contrast, does hold that at the fundamental level, the most basic things do possess non-physical properties).

[9]There are of course other problematic domains: mathematical and other abstract entities, aesthetic and moral properties are examples. In all these cases, however, it is not altogether implausible to suggest that either they do not exist at all or are in some way reducible to the responses of conscious subjects. Consciousness itself still stands apart.

[10]For an overview of the problematic dialectical situation of physicalism see Seager 2016, ch. 1.

[11]The pandemonium model traces back to early work in cognitive science and computer engineering; see Selfridge (1959). In a much more disciplined and structured form than in Dennett's models the pandemonium idea remains a core idea in computer system design.

[12]Sometimes the demons are likened to politicians, but sometimes to cultural celebrities. Celebrities are widely known, emulated, deferred to, but only so long as their fame lasts. Hence Dennett's explications of consciousness as 'cerebral celebrity' or 'fame in the brain'.

[13]Presumably, evolutionary pressures will tend to lead to organisms that by and large believe the true and desire what is good for them (see Dennett 1991a, ch. 7; 1978b; Seager 2000) and hence to organisms to which the intentional stance can be successfully applied.

[14]It is relevant here that Ryle was Dennett's doctoral supervisor.

[15]An interesting exploration of the radical nature of Dennett's pattern metaphysics can be found in Haugeland (1998). Dennett's views can be usefully compared to Nicholas Rescher's 'conceptual idealism' (1991).

[16]Dennett's arguments are highly contentious in any case; for an assessment see Seager (2016), chs. 7,8.

[17]Jackson's argument is by now the subject of very large literature; see Ludlow *et al.* (2004) for an excellent collection of articles devoted to its analysis and criticism.

[18]Here I use some ideas from Nida-Rümelin (204).

[19]It is deeply interesting that the spatial case seems so obviously devoid of novel information compared to the temporal case. Experience seems to intrinsically suggest that we all share the same *now*, even as this seems contrary to modern physical understanding.

References

Aguirre, Geoffrey K. and Mark D'Esposito (1999). 'Topographical disorientation: a synthesis and taxonomy'. *Brain*, 122 (9): pp. 1613–1628.

Bianchini, F., C. Incoccia *et al.* (2010). 'Developmental topographical disorientation in a healthy subject'. *Neuropsychologia*, 48 (6): pp. 1563–73.

Cao, Tian and Silvan Schweber (1993). 'The Conceptual Foundations and Philosophical Aspects of Renormalization Theory'. *Synthese*, 97: pp. 33–108.

Carroll, Sean (2010). 'The laws underlying the physics of everyday life are completely understood'. URL `http://blogs.discovermagazine.com/cosmicvariance/2010/09/23/the-laws-underlying-the-physics-of-everyday-life-are-completely-understood/`. *Cosmic Variance*, September 23.

Castellani, Elani (2002). 'Reductionism, Emergence and Effective Field Theories'. *Studies in History and Philosophy of Modern Physics*, 33 (2): pp. 251–67.

Chalmers, David (2009). 'The Two-Dimensional Argument Against Materialism'. In B. McLaughlin, A. Beckermann and S. Walter (eds.), *The Oxford Handbook of Philosophy of Mind*, pp. 313–38. Oxford: Oxford University Press.

Churchland, Paul (1985). 'Reduction, Qualia, and the Direct Introspection of Brain States'. *Journal of Philosophy*, 82: pp. 8–28.

Dennett, Daniel (1971). 'Intentional Systems'. *Journal of Philosophy*, 68 (4): pp. 87–106. Reprinted in Dennett's *Brainstorms*, Cambridge: MA, Bradford Books, 1978.

Dennett, Daniel (1978a). 'Two Approaches to Mental Images'. In *Brainstorms: Philosophical Essays on Mind and Psychology*, pp. 174–89. Montgomery: VT: Bradford Books.

Dennett, Daniel (1978b). 'Why the Law of Effect Will Not Go Away'. In *Brainstorms: Philosophical Essays on Mind and Psychology*. Montgomery: VT: Bradford Books. Originally published in *The Journal of Theory of Social Behavior*, 2, pp. 169-87.

Dennett, Daniel (1988). 'Quining Qualia'. In A. Marcel and E. Bisiach (eds.), *Consciousness in Contemporary Science*. Oxford: Oxford University Press. (Reprinted in Lycan and Prinz (2008)).

Dennett, Daniel (1991a). *Consciousness Explained*. Boston: Little, Brown & Co.

Dennett, Daniel (1991b). 'Real Patterns'. *Journal of Philosophy*, 88: pp. 27–51. Reprinted in Dennett's *Brainchildren: Essays on Designing Minds*, Cambridge, MA: MIT Press, 1998.

Dennett, Daniel (2001a). 'Are We Explaining Consciousness Yet'. *Cognition*, 79: pp. 221–37.

Dennett, Daniel (2001b). 'Consciousness—How Much is that in Real Money?' In R. Gregory (ed.), *Oxford Companion to the Mind*. Oxford: Oxford University Press. Reprinted in Dennett (2005).

Dennett, Daniel (2002). 'Brian Cantwell Smith on Evolution, Objectivity and Intentionality'. In Hugh Clapin (ed.), *Philosophy of Mental Representation*, pp. 222–36. Oxford: Oxford University Press.

Dennett, Daniel (2005). *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*. Cambridge, MA: MIT Press (Bradford Books). (2005 Jean Nicod Lectures).

Dennett, Daniel and Kinsbourne Marcel (1992). 'Time and the Observer'. *Behavioral and Brain Sciences*, 15 (2): pp. 183–201.

Feynman, Richard (1985). *QED: The Strange Theory of Light and Matter*. Princeton: Princeton University Press.

Fine, Kit (2008). 'Coincidence and Form'. *Aristotelian Society Supplementary Volume*, 82 (1): pp. 101–18.

Haugeland, John (1998). 'Pattern and Being'. In *Having Thought: Essays in the Metaphysics of Mind*, pp. 267–90. Cambridge, MA: Harvard University Press.

Heil, John (2003). *From an Ontological Point of View*. Oxford: Oxford University Press.

Jackson, Frank (1982). 'Epiphenomenal Qualia'. *Philosophical Quarterly*, 32: pp. 127–36.

James, William (1887). 'The Perception of Space'. *Mind*, 12 (45): pp. 1–30. Reprinted with revisions in James's *The Principles of Psychology*, 1890, Henry Holt and Co: New York, ch. 20.

James, William (1904a). 'A World of Pure Experience'. *Journal of Philosophy, Psychology and Scientific Methods*, 1: pp. 533–43, 561–70.

James, William (1904b). 'Does "Consciousness" Exist?' *Journal of Philosophy, Psychology, and Scientific Methods*, 1: pp. 477–91.

Kant, Immanuel (1786/1998). 'What Does it Mean to Orient Oneself in Thinking'. In Allen Wood and George Di Giovanni (eds.), *Kant: Religion Within the Boundaries of Mere Reason: And Other Writings*, pp. 3–14. Cambridge: Cambridge University Press.

Locke, John (1690/1975). *An Essay Concerning Human Understanding*. Oxford: Oxford University Press (Clarendon).

Ludlow, P., Y. Nagasawa *et al.* (eds.) (2004). *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. Cambridge, MA: MIT Press.

Lycan, William and J. Prinz (eds.) (2008). *Mind and Cognition: An Anthology, 3rd Edition*. Oxford: Blackwell.

McLaughlin, Brian (1992). 'The Rise and Fall of British Emergentism'. In A. Beckermann, H. Flohr and J. Kim (eds.), *Emergence or Reduction*, pp. 49–93. Berlin: De Gruyter.

Nagel, Thomas (1974). 'What Is It Like to be a Bat?' *Philosophical Review*, 83 (4): pp. 435–50. (This article is reprinted in many places, notably in Nagel's *Mortal Questions*, Cambridge: Cambridge University Press, 1979.).

Nida-Rümelin, Martine (204). 'What Mary Couldn't Know: Belief About Phenomenal States'. In P. Ludlow, Y. Nagasawa and D. Stoljar (eds.), *There's Something About Mary*, pp. 241–68. Cambridge, MA: MIT Press.

Perry, John (1977). 'Frege on Demonstratives'. *The Philosophical Review*, 86 (4): pp. 474–97.

Perry, John (1979). 'The Problem of the Essential Indexical'. *Nous*, 13 (1): pp. 3–21.

Perry, John (2001). *Knowledge, Possibility and Consciousness*. Cambridge, MA: MIT Press.

Pitt, David (2004). 'The Phenomenology of Cognition, Or, What Is It Like to Think That P?' *Philosophy and Phenomenological Research*, 69: pp. 1–36.

Rescher, Nicholas (1991). 'Conceptual Idealism Revisited'. *The Review of Metaphysics*, 44 (3): pp. 495–523.

Ryle, Gilbert (1949). *The Concept of Mind*. London: Hutchinson & Co.

Seager, William (2000). 'Real Patterns and Surface Metaphysics'. In Don Ross, Andrew Brook and David Thompson (eds.), *Dennett's Philosophy: A Comprehensive Assessment*, pp. 95–130. Cambridge, MA: MIT Press.

Seager, William (2016). *Theories of Consciousness*. London: Routledge, 2nd ed.

Selfridge, Oliver (1959). 'Pandemonium: A Paradigm for Learning'. In *Mechanization of Thought Processes*. London: Her Majesty's Stationery Office. (Reprinted in P. Dodwell (ed.) *Perceptual Learning and Adaptation* (1970), Hammondsworth: Penguin, pp. 465ff.).

Shoemaker, Sidney (1980). 'Causality and Properties'. In P. van Inwagen (ed.), *Time and Cause: Essays Presented to Richard Taylor*, pp. 109–35. Dordrecht: Reidel. (Reprinted in Shoemaker's *Identity, Cause and Mind*, Oxford University Press: Oxford, 2003, pp. 206-33).

Stalnaker, Robert C (2008). *Our Knowledge of the Internal World*. Oxford: Oxford University Press.

Suppe, Frederick (1977). 'The Search for Philosophic Understanding of Scientific Theories'. In Frederick Suppe (ed.), *The Structure of Scientific Theories*, pp. 3–241. Urbana: University of Illinois Press, 2^{nd} ed.

Webb, J. K., J. A. King *et al.* (2011). 'Indications of a Spatial Variation of the Fine Structure Constant'. *Phys. Rev. Lett.*, 107: p. 191101.