Alien Subjectivity

and the Importance of Consciousness[1]

Geoffrey Lee

1 : Introduction

Why are we interested in studying consciousness[2]? One reason is that it is peculiarly difficult to understand. There is no consensus over what basic type of thing it is (is it a special kind of representation of the world? A kind of self-representation? A primitive contact with sense-data?). More fundamentally, it is mysterious to us how it relates to physical phenomena – it seems very different from say, a complex pattern of neural firing. There is therefore an interesting project of figuring out what it is, and how it fits into the natural order of things (this is, of course, a major part of the mind/body problem).

A second reason why we are interested in it though, is that we think of it as objectively significant, in several senses. I would distinguish a *descriptive, epistemic* and *moral* sense in which it seems significant (the first may encompass the others). It seems descriptively significant, in the sense that if someone were to give a characterization of the universe we live in – in particular, that region of space-time occupied by humans and other sentient life forms – they would not have adequately described it if they didn't mention the presence of consciousness: consciousness seems like a deep joint in nature, or a highly natural property. It seems epistemically significant in the sense of being a special way of getting epistemically connected to things. Conscious perceptual acquaintance with an event can justify beliefs about the event, and seems to be special kind of epistemic state in its own right. Furthermore, conscious mental events themselves seem to be epistemically accessible in a peculiarly intimate way: consciousness seems to be self-illuminating. Finally, consciousness appears to be a source of much of what is valuable in the universe. Retaining consciousness is necessary for retaining what is valuable in living, and beings that lack consciousness would appear to lack something necessary for treating them as intrinsically worthy of moral respect.

---

[1] Thanks are due to Tony Beszylko, John Campbell, David Chalmers, Uriah Kriegel, Mike Martin, Carlos Montemayor, Laurie Paul, Jon Simon, Declan Smithies, James Stazicker, Brad Thompson, and participants at the 2010 New Directions in Philosophy of Mind Workshop at Columbia University for helpful comments and discussion. Thanks guys!

[2] To be clear, the kind of "consciousness" I'm interested in here is *phenomenal* consciousness, in the sense of there being "something its like" for a subject- as opposed to self-consciousness, or the perceptual sensitivity that comes with being awake and alert. Some philosophers doubt whether there is a clear sense attached to "phenomenal consciousness" – a view that I consider a form of Deflationism to be contrasted with the Deflationary view I conditionally defend in this paper.

Strawson (1994) suggests something like this sentiment in the following passage:

> It is true that the line between mental or experiencing beings and others may look unimportant from the point of view of animal ethology and general biology, which study the behavior of all living organisms without any regard to experience. The fact remains that it is a line of great importance. It is arguably the most important theoretical line to be drawn in the whole of reality. (Strawson (1994) p.154, quoted in Smithies (ms.))

It is this idea that consciousness is highly significant in various respects that I want to focus on. Can we clarify the sense (or senses) in which consciousness is supposed to be a theoretically "deep" distinction? And is it really true that it is a distinction with this kind of depth? If property dualism is true, then consciousness is a fundamental ingredient in reality (or more determinate experiential properties are), and so has obvious significance in at least one sense[3]. But what if instead consciousness is a complex high-level physical or functional property of the brain or some larger physical system? One might reasonably wonder whether this is consistent with its having the kind of significance Strawson is gesturing at in the quoted passage. The main aim of this paper is to develop this deflationary thought, by arguing that there is a kind of deflationary stance towards the significance of consciousness, which I call *deflationary pluralism*, which is plausible if reductive materialism is correct. When we focus on the possible consequences of this view, we'll see that reductive materialism may commit one to a more deflationary attitude towards consciousness than many reductive materialists, such as Ned Block, may have thought they needed to accept.

According to the deflationary pluralist, there is no single deep distinction between conscious beings and the rest, but rather a family of "consciousness-like" properties that are about equally significance in a given respect (descriptively, epistemically, morally). At least two surprising consequences can be argued to follow from this. First, that a completely unconscious creature – a complex alien or artificial life form, perhaps – could have a state that played a similar role in its mental life to consciousness, and which is equally significant as consciousness (in a certain respect, perhaps). Intuitively, the idea is that there is no special "glow" attached to all and only conscious beings; a zombie could have a quasi-conscious internal state that is just as glowingly special as consciousness. Second, the question of whether or not to attribute consciousness in certain problem cases, such as the cases of lobsters, or certain intelligent robots or aliens, or in the case of certain human psychological states, might turn out to be an empty question. For example, assuming we know everything about how a problematic creature physically functions, there is no further substantive fact to be learned by finding out whether they are conscious (more below on what this means).

Both the view and these alleged consequences have helpful analogies in the view of Personal Identity advocated by Derek Parfit (1984, 1995) (although there are

---

[3] Similar remarks apply in the cases of panpsychism and Russelian monism.

disanalogies too[4]). In the case of personal identity, we are inclined to think of the difference between life and death as a deep gulf, the gulf between our still having a subjective viewpoint on the world, and the "light inside the skull" being permanently switched off – total, unimaginable, permanent nothingness. Similarly, when thinking about consciousness there appears to be a deep distinction between those creatures who are such that "the lights are on inside" – there is something its like for them, they have a subjective view – and those for whom the unimaginable darkness of not experiencing the world – of not being conscious - is the norm. However, on Parfit's view, the underlying facts in virtue of which personal identity obtains if a reductionist view of persons is correct – certain physical or psychological relations between person stages – are just not of the right kind to make this picture sustainable. In particular, according to Parfit, there can be cases where although you no longer exist, you have some *just as good* as continuing to exist. Similarly, I want to suggest that if reductionism about consciousness is correct, then there is no deep gulf between conscious beings and the rest, which arguably means that being unconscious can sometimes be just as "good" as being conscious. Roughly, this is the deflationary pluralist's position.

Parfit also thinks that in some cases asking whether one will still exist in a given scenario – for example, in a case of teletransportation – is an empty question. The question is empty, not in the sense that it has no answer, but in the sense that learning the answer won't tell you any interesting information about the world – more specifically, it won't tell you *where a significant boundary lies* (see Sider (2011) chapter 3, for a helpful development of this conception of an empty question. In effect, Sider's view is that there will empty or "non-substantive" questions concerning the application of a concept exactly when a form deflationary pluralism is true for the concept; I am therefore considering the claim that some questions about consciousness are non-substantive in Sider's sense. Chalmers (2009) discussion of merely verbal questions is also relevant here). Interestingly, Parfit explicitly considers and rejects an analogous view for consciousness:

> ….Suppose we are studying some creature which is very unlike ourselves, such as an insect, or some extra terrestrial being. We know all the facts about this creature's behavior, and its neurophysiology. The creature wriggles vigorously, in what seems to be a response to some injury. We ask, "Is it conscious, and in great pain? Or is it merely an insentient machine?" Some Behaviourist might say, "That is a merely

---

[4] One disanalogy is that Parfit is exclusively interested in the practical or moral significance of personal identity, whereas my deflationary pluralist might be concerned with the descriptive or epistemic significance of consciousness, as well as its practical or moral significance. A second important disanalogy is this: although Parfit thinks that personal identity isn't as significant as we ordinarily think, on his view there is *another* relation – the relation of psychological continuity and connectedness – that *is* what's really significant. By contrast, it is not part of the Deflationary Pluralist's view of consciousness that there's some property other than consciousness that we ought to regard as outstandingly significant instead.

> verbal question. These aren't different possibilities, either of which might be true. They are merely different descriptions of the very same state of affairs." That I find incredible. These descriptions give us, I believe, two quite different possibilities. It could not be an empty or merely verbal question whether some creature was unconscious or in great pain. (Parfit (1995), p. 26)

I think that many of those who are reductive materialists about consciousness would still be inclined to agree with the sentiment Parfit is expressing in this passage. According to the Deflationary Pluralist, this is a mistake: no complex material property could constitute the gulf in nature that this would require.

I should stress that my argument for Deflationary Pluralism is very much conditional on the truth of reductive materialism. I take reductive materialism to be the view that phenomenal consciousness and more determinate phenomenal properties – the properties whose instantiation constitutes what it's like for a subject - have complex real definitions in non-psychological terms. This will imply that everything about consciousness obtains in virtue of the distribution of non-mental properties, which are thereby more fundamental than phenomenal properties. I will have more to say about the notions of "fundamentality" and "natural property" that are needed to make sense of these ideas below. For now I should stress that these real definitions need not be available a priori, or even on the basis of a complete specification in physical terms of which centered possible world you are in. The definitions could be infinitary (e.g. infinite disjunctions), although I think that is implausible. They could also be in functional or structural terms, involving second-order quantification over determinate physical quantities like mass, charge and spin: in fact I think this is probably the most plausible version of the reductionist position.

In my opinion, *non*-reductive views are best understood in terms of an ideology of "perfectly natural" properties, as defended by Armstrong and Lewis (see Armstrong (1978), Lewis (1983)), among others (again, I will say more about this below). We can understand anti-reductionism as the view that some phenomenal properties are perfectly natural properties. For my purposes, it will be best to understand anti-reductionism as consistent with the view that phenomenal events supervene on non-phenomenal events, and are therefore not "fundamental" in the sense that they are part of a base of perfectly natural properties whose instantiation determines everything else, and whose instantiation is not itself determined from below. For my purposes, the important question will be whether phenomenal properties are primitive non-reducible properties, *not* the further question of whether they are also fundamental in this sense.

An outline of what is to come: in section 2, I compare deflationary pluralism with a number of other deflationary stances one could take, and say more about what the view is. In section 3, I make a distinction between two different views of how best to *ascertain* the significance (either descriptive, epistemic, or moral) of a property – a bottom up view, and a top down view. My basic argument from reductive materialism to deflationary pluralism will be that Reductive Materialism supports a

bottom up epistemology of significance, which in turn supports Deflationary Pluralism. In section 4 I defend a version of the argument that focuses exclusively on Descriptive Significance, including a discussion of what is involved taking a property to be descriptively significant (which is supposed to be cognate with "natural" or "joint-carving", in the sense discussed by Armstrong (1978) and Lewis (1993)). Finally, in section 5, I briefly outline some connections with other issues and some questions that need further discussion.

The epistemic significance of consciousness, and the moral or practical significance of consciousness, despite their obvious importance, won't be pursued here, except for some brief remarks about the epistemic significance of consciousness in section 4. Arguably there is a rationale for focusing first on the naturalness of consciousness before looking at its normative significance, which is that we normally think of these normative kinds of significance as attaching to consciousness in a way that *depends* on its natural significance. For example, consider the difference between real pain and unconscious functional analogue of pain occurring in a zombie. We think that zombie pain isn't bad in the way that real pain is, *because* its nothing like real pain. That is, the big difference between real pain and zombie pain doesn't seem to *consist* in the fact that one is bad and the other isn't, rather it seems that there is a big natural difference between them that *grounds* the fact that one is bad and the other isn't. One could imagine a view that denies this, but I think it would be revisionary of normal thinking on the matter. (see Lee (manuscript) for a much more extensive discussion of the epistemic and practical significance of consciousness, and their connection with its natural significance).

Before proceeding, an important point about this discussion is worth immediately emphasizing: it would be a mistake to equate being deflationary about the existence of a so-called "explanatory gap" with being a deflationary pluralist, or being a deflationist in any of the other senses I'll discuss. For example, one could be a realist about the explanatory gap, but nonetheless think that from a cosmic perspective consciousness is really nothing special; its interest lies entirely in the explanatory puzzle it poses *for us* (Mcginn (1989) might be read as expressing this view). In the same spirit, perhaps we can't ever know whether lobsters or certain intelligent computers are conscious; it doesn't follow that we lack any *substantive knowledge*, in the sense of knowledge of where a significant boundary lies. On the flipside, one could deny that there's a gap (e.g. by holding an analytical functionalist view like Lewis' or Shoemaker's (see Lewis (1972), Shoemaker (1975)), but still agree with Strawson's grandiose sentiment. Despite this kind of independence, there are some interesting connections to be drawn here that I'll briefly describe at the end of the paper. I'll be generally assuming throughout that there *is* an explanatory gap, although for the reasons just given, not much turns on this.

2: Deflationary Pluralism vs Other Deflationary Approaches

Deflationary Pluralism is not the only view that challenges the thought that consciousness marks a deep distinction in nature. In this section I distinguish a

number of positions of this kind, and say more about what is distinctive about the DP position.

One way to argue that there is no deep distinction between conscious beings and the rest would be to argue that there *are no* conscious beings : one could hold an error theory. I think most philosophers would reject this view, taking the existence of conscious experiences as a Moorean fact. I won't seriously challenge this starting point here.

Similar in metaphysical outlook to the error theory, there is the expressivist view that statements about experience are not genuinely descriptive or fact-stating but really express a non-cognitive attitude like an evaluation. This view is implausible, despite the fact that we think of consciousness as highly normatively significant in both an epistemic and moral way. Although we think that being conscious gives you a special kind of epistemic access to your own states and to the environment, and we think that it gives you a special moral status, we don't think that saying that an individual is conscious is just the *same thing* as making an epistemic or moral evaluation of them. Consciousness seems like a robust part of the world.

One could argue that consciousness isn't significant in a certain respect by arguing that *nothing* is significant in that respect – that is, by holding a *nihilist* view of a certain kind. One could be a sceptic about the existence of objective moral or epistemic norms, or one could be a sceptic about the idea that some distinctions are more descriptively significant than others, i.e. that some properties are more "natural" or "joint-carving" than others. I won't be trying to further clarify or assess Nihilism here, except in the case of descriptive significance, which I return to below.

Even if one is not a nihilist of any sort, one might hold the view that consciousness positively lacks significance of a certain sort, even though *other* properties are relatively significant in the same sense. For example, one might hold that consciousness is a *disjunctive* property, and therefore has a low degree of naturalness – there is nothing deeply in common between all the instances of phenomenal consciousness (this makes sense if we are realists about naturalness: a disjunctive property is a disjunction of relatively natural properties). Or one might hold that consciousness is *epiphenomenal*, and therefore has a much lower degree of causal-explanatory significance than other properties. Or one might hold that although we have beliefs or other mental states that are objectively justified (or otherwise have positive epistemic status) this has nothing, or much less, to do with conscious experience of the world than we ordinarily think.

Ned Block (2002) takes seriously an inter-personal version of the disjunctive view, holding that we can't rule out the view that his Android Commander Data is phenomenally conscious, despite the fact there is no natural property he shares with us in virtue of which he is phenomenally conscious. Alternatively, one could consider an intra-personal disjunctive view, according to which different phenomenal states within a single individual need not have anything deeply in

common. I think both views (especially the inter-personal view) are implausible, and they are not my main concern here (see footnote for some brief discussion)[5].

Note that the claim that consciousness is highly natural property, and the claim that it has great causal-explanatory significance are conceptually distinct, as is illustrated by the coherence of epiphenomenalist dualism: phenomenal properties could be a fundamental part of the world, even if they do no causal work at all. Nonetheless, later on I will appeal to a view on which naturalness and causal-explanatory significance are strongly linked.

Fred Dretske (2000) thinks that it's a no-brainer that conscious states have great causal-explanatory significance, at least with respect to our behavior, on the grounds that if we weren't conscious we would be blind and bump into things all the time. Perhaps this is fair comment, but it fails to engage with what is perhaps the more interesting issue in the vicinity: could there be an unconscious creature that had a kind of internal state that did the work conscious states do in us, but without thereby making the creature conscious? Is consciousness *unique* in the kind of significance it has (in this case, causal-explanatory significance)?

This brings us to Deflationary Pluralism. Not only do we think that consciousness has descriptive, epistemic and moral significance, but we also think that it has *uniquely* strong significance in each of these ways – it stands out from the crowd. It is this stronger claim that the Deflationary Pluralist challenges: it is not that consciousness is positively unnatural, or epiphenomenal, or otherwise insignificant; it is just not uniquely or outstandingly deep. I will now explain this view in more detail.

It will be helpful to introduce some terminology. Let's say that a property is *consciousness-like* if it has similar functional role to consciousness as it is found in humans. Similarly, for more specific types of conscious experience, we can talk about properties that are e.g. pain-like, color-perception-like and so on. I'll say that a being is *pseudo-conscious* if they are in a consciousness-like state. There are various

---

[5] Two objections to the inter-personal disjunctive view : first, it comes uncomfortably close to the view that really there is no such thing as consciousness. If we discover that some apparently unified category is disjunctive, it is arguably just a verbal matter whether we say that the original property is instantiated, although it is disjunctive, or whether we say instead that there is no such thing. For example, it would not be unreasonable, on discovering that "jade" does not cover a uniform category of things, to say that there is no such thing as jade, just these more specific kinds of jewels. Second: it is a plausible meta-semantic assumption that our terms refer to non-disjunctive natural kinds, unless there is considerable pressure from the facts of usage to interpret them disjunctively (e.g. we are deliberately trying to pick out a disjunctive kind). But there is nothing about our use of "conscious" that presses us to interpret it as picking out a disjunctive property that includes Data in its extension.

ways of fleshing the kinds of functional similarities required in order to be in a consciousness-like state. We could require duplication of only functional features known to the folk to be associated with consciousness; or we could require duplication of functional features known empirically to be associated with conscious experience. Thus we have a distinction between properties that are *folk-consciousness-like* and *empirically-consciousness-like*. We could also trade only in functional constraints that are necessarily satisfied by any conscious being, or also include constraints that are at best only contingently satisfied in conscious humans.

Now let's say that a being is *quasi-conscious* if they have a property that is consciousness-like *and* is equally significant as consciousness. We can make this specific to a particular kind of significance. For example, we can talk about a being that is descriptively, epistemically, or morally quasi-conscious. By analogy, we can also talk about quasi-pain, quasi- visual experience, and so on. I think the common-sense view is that necessarily, all and only quasi-conscious individuals are conscious. What unites various forms of Deflationary Pluralism is that they hold that there are ways to be quasi-conscious without being conscious. Beyond this, we get different forms of DP depending on which consciousness-like properties are held to be sufficient for quasi-consciousness, and which kind of quasi-consciousness is at issue: descriptive, epistemic, or moral. It'll be useful to distinguish views according to whether they hold that some or all of the following consciousness-like properties are sufficient for quasi-consciousness:

(1) Small variations: Properties that differ only slightly from "Phenomenal Consciousness" in where they draw their boundaries in a given actual or possible situation.
(2) Properties at different levels : Functional Properties that are either richer than, or more abstract than the property of satisfying the functional constraints associated with being conscious in humans, corresponding to different levels of functional analysis from the level where those constraints can be specified.
(3) Alien Realizations of type (2) properties : Alien physical realizations of the superficial functional architecture associated with consciousness (i.e. properties that *overlap* with consciousness in its functional role).

To illustrate a type (1) view, let us suppose that lobsters are conscious, but they fall just barely within the extension of our concept "conscious". We can imagine a different concept that is used very much like ours (i.e. it has a similar conceptual role) except that it draws the boundary in a slightly different place, so that lobsters, as they actually are, are not included in the set of possible objects the term applies to. We can imagine a group of individuals, the "lobster exclusionists", who use this concept rather than the concept of consciousness.

I think common sense would say that there is something defective about their concept. For example, suppose that both us and the exclusionists have a belief that we would express with the sentence "if lobsters feel conscious pain, it is wrong to boil them alive". Since we mean something different by "conscious", it may be that

their conditional delivers the result that it's ok to boil lobsters, even though ours gives the opposite result. Common sense says they are making a mistake. Their concepts don't pick out the phenomenon that really is morally significant, so they are incapable of conceptualizing the important similarity between us and lobsters that dictates the appropriateness of a certain attitude towards them.

One kind of Deflationary Pluralist (type 1) disagrees. According to them, the property picked out by their concept of "consciousness" (or "pain") is just as significant (perhaps in a given respect, e.g. explanatorily, epistemically, morally) as the property picked out by ours. So they aren't missing an important distinction, and their conceptual practice is just as reasonable as ours.

We can illustrate type (2) and (3) DP views by considering beings that are *mere superficial duplicates* of us in the sense that they enjoy a consciousness-like functional property, but beyond that they have nothing in common with us. Different kinds of consciousness-like properties generate different kinds of mere superficial duplication. A mere folk functional duplicate has internal states that play causal roles corresponding to common sense platitudes about the causal roles of folk-psychological states like beliefs and desires; but this architecture is realized in them in completely different way from in us (perhaps subject to certain constraints[6]). This is the kind of architecture that Block discusses in his (2002), as enjoyed by the android Commander Data. By contrast, a mere *empirical* duplicate has a psychological architecture superficially like the one that we discover through empirical research is associated with consciousness, but is otherwise as different from us as possible. For example, we can imagine a creature that has relatively modularized perceptual processing that feeds information via a selective mechanism into a short-term memory store that is then either placed in long-term storage or used in other information processing tasks that lead to actions. But beyond this kind of high-level similarity, the architecture is realized in a completely different way in us.

If we reject the view that each kind of superficial duplication is sufficient for consciousness (and we reject the intra-personal disjunctive view canvassed above (see footnote[7])), then we can conclude that an alien or robot that merely duplicated

---

[6] In his (2002) discussion Block limits the extension of "folk functional duplicate" in a way that excludes various odd cases where we have the intuition that the being in question isn't conscious – for example, the case where the folk-functional architecture is realized by a head full of little homunculus creatures communicating with each other. For my purposes the important issue is whether a homunculus headed creature is quasi-conscious but not conscious. A proponent of a type 2 view who has the intuition that the homunculus head is neither conscious *nor* quasi-conscious may wish to limit the extension of "folk-functional duplicate" in a way similar to Block, so that having such an architecture really is sufficient for being quasi-conscious.

[7] Some theorists, such as Block, take seriously the thought that although Superficialism is false, consciousness might nonetheless be disjunctively realized in a way that means our alien or robot is conscious. For the reductionist, this means taking consciousness to be a

us in one of these ways would *not* be phenomenally conscious. I'll return below to whether this rejection of a "Superficialist" view of consciousness is plausible. For now, let's consider how we would ordinarily think about these aliens or robots, on the assumption that they *aren't* conscious. Common-sense says that us and the aliens are on either side of deep gulf in nature, lacking a property that is extremely significant in various ways. For example, because all is dark within, their mental lives are massively epistemically impoverished, and they fail to have the moral status of human beings. A type 2 or 3 Deflationary Pluralist disagrees. According to them, at least some such unconscious aliens still have a consciousness-like property that is equally significant as consciousness. For example, suppose that consciousness (*our* "consciousness") involves a distinctive kind of globally synchronized 40 hz neural firing that implements a special kind of integration of information from different cognitive modules. The aliens don't have this, but they may have an equally interesting property that implements the superficial role of consciousness in them (an instance of a type 3 view). Perhaps they don't have anything like neurons, and so although they have informational integration, it's not achieved by anything like synchronized firing. So (let's assume) they aren't conscious : we can't say there's anything it's like for them. But maybe there is something it's *schlike* for them!

We can illustrate the difference between type (2) view and a type (3) view as follows. A type (2) theorist is a pluralist about the significance of different consciousness-like properties that humans actually instantiate. The idea is that there is no single property or level of description of our cognition that pops out as special, the level where consciousness resides. By contrast, a type (3) theorist focuses on alien properties that we *don't* instantiate. It is consistent with their deflationary view that there really is a single special consciousness-like property that stands out relative to the other properties *we* instantiate. Perhaps consciousness really *is* special relative to anything else we know about (it is "locally outstanding"). It's just that an alien could enjoy a *different* consciousness-like property that stands out *just as much*. Their quasi-consciousness is just as deep a joint in nature, or is just as epistemically or morally significant as conscious states are for us. Among consciousness-like properties, consciousness might be locally, but not globally outstanding.

These are the kinds of "consciousness-like" properties I want to focus on (there may be viable versions of DP which involves other kinds). A type (1) DP view that limits itself to "small variations" on consciousness ought to be the least controversial version of the view. How could a property that is only very slightly different from consciousness be much less significant than consciousness? I imagine that on reflection, most materialists will accept this position, even if in some ways it conflicts with common sense. Therefore, I'll focus in what follows on type 2 and 3

---

disjunctive (and therefore highly unnatural) property, a view I rejected for reasons given in footnote 4.

versions of the view, which may seem more radically revisionary of common sense than a mere commitment to a reductive materialist view would seem to entail.

It's important to note that a Deflationary Pluralist needn't hold the view that there is some natural property, such as a particular consciousness-like functional organization, that all and only quasi-conscious beings have in common (at least, not beyond the 2nd order property of having a significant consciousness-like property). The only 1st order property in common to all possible quasi-conscious beings might be a disjunction of consciousness-like properties (this is similar to the situation we have with the 2nd order property of "instantiating a fundamental property"). This fact is important to bear in mind when assessing the *Superficialist response* to DP, which says that all the allegedly unconscious yet quasi-conscious aliens we are considering are really conscious after all – their superficial functional organization is sufficient for consciousness. That is, consciousness is a sufficiently broad kind that it includes all instances of quasi-consciousness.

Against this, consider a version of DP on which there is a variety of different consciousness-like properties that are as significant (in some respect) as consciousness. Assuming that consciousness isn't a disjunctive property, the proponent of the Superficialist response will have to hold that there is a natural property sufficient for consciousness that is shared by all the corresponding quasi-conscious beings (and which may belong to other creatures as well). The trouble is that this is might have to be an *extremely* abstract functional property, e.g. more abstract than a property that is sufficient for folk-functional duplication. For example, following Mclaughlin (2003), consider the fact that we are instinctively confident that a human baby is conscious, despite the fact that they are not a folk-functional duplicate of a human adult. Consider a Deflationary Pluralist view on which both Block's Commander Data and such a human baby are quasi-conscious. They might have nothing more than an extremely abstract functional architecture in common, perhaps an architecture that they share with a bumble bee or a smartphone. So, holding they are both conscious seems to lead to *radical superficialism*.

Against this view, we might think it implausible that the bumble bee has consciousness, or at least think that if it does, consciousness is such an abstract property that we have an independent reason to doubt whether it really is significant. Compare this with a Deflationary Pluralist view on which, although one out of the baby or Data is not conscious, they both have different consciousness-like internal states that are as significant as consciousness. That is, they are quasi-conscious in virtue of *different* resemblances with an adult human. This strikes me as a much more attractive view than the corresponding radical form of

Superficialism. (A different way to reply to the Superficialist is to argue that the view isn't really inconsistent with Deflationary Pluralism (see footnote)[8].)

I mentioned earlier than if DP is true, then asking whether a creature is conscious in some problem case could turn out to be asking an empty question: a question whose answer doesn't tell us where a significant boundary lies. The reason why the question whether a creature is conscious seems so significant is that we think lacking consciousness means *lacking* a state with the kind of significance that consciousness has. But if lacking consciousness is consistent with having a form of quasi-consciousness that "glows" just as much as consciousness does, then the boundary between conscious beings and the rest isn't as significant as we ordinarily think. Hence, learning that a being isn't conscious isn't necessarily as deeply informative as it might seem; indeed it may not be informative at all (more on this below). In this sense, DP could imply that some questions about the boundaries of consciousness are empty.

Many readers might reasonably be wondering how the claim that some debates about consciousness are empty or insubstantive relates to the claim that certain propositions about consciousness aren't determinately true, or that the corresponding questions have no answer. Before proceeding, it'll be worth making a few brief remarks about this.

The claim that "phenomenal consciousness" is vague or indeterminate is different from the uncontroversial claim that "consciousness" is an ambiguous term in English (here, as I hope is clear, I am intending to use it to refer to *phenomenal consciousness,* rather than alertness, responsiveness or self-awareness, to mention a few other candidate disambiguations). There is a strong intuition that many people have that it can't be indeterminate whether there is something its like to be some creature. Either the lights are turned on inside or they aren't (fading or faint consciousness isn't a relevant example, because it's still determinately a form of consciousness). Furthermore, I think it is probably true that this feeling of a *sharp* divide is at least a factor in explaining why consciousness seems like a *deep* divide.

We have good reason to be skeptical of these intuitions if we are reductive materialists. For the RM, the presence of consciousness presumably depends on the presence of a number of continuously variable physical magnitudes, meaning that the location of any sharp boundary for consciousness will be highly arbitrary. Furthermore, the fact that we can't conceive "from the inside" of being only

---

[8] Suppose that a suitably radical form of Superficialism is true. There could still be a class of "consciousness-like" states that are equally significant as consciousness; its just that having one of these states will require being conscious: that is, quasi-consciousness will require consciousness. Still, if we imagine an impartial observer comparing us with an alien that merely superficially duplicates us, they probably wouldn't see the properties we share in common as greatly more significant than those that differ between us. In this way, even in the context of Radical Superficialism, we can still argue for something like a Deflationary Pluralist view.

indeterminately conscious might be no better evidence that this is impossible, than the fact that we can't conceive of being *not* conscious "from the inside" is evidence that *that* is impossible. In both cases, this inconceivability is probably telling us more about the nature of "imagination from the inside" than the nature of consciousness itself (having said this, there are considerations that suggest that indeterminate consciousness a very unusual case of indeterminacy[9]).

The Deflationary Pluralist view is closely related to the view that "phenomenal consciousness" is indeterminate, but it is importantly different. On certain meta-semantic assumptions, such as Lewis's theory that reference is determined by a trade off between fit with usage and naturalness, if we enjoy a single highly natural consciousness-like property, then this property will act as a "semantic magnet" for our term "consciousness", and it will therefore almost certainly have a determinate referent. But if Descriptive Deflationary Pluralism is true, in particular if a "local" (type 1 or type 2) form of the view is true, then there may be no such outstanding candidate, and therefore indeterminacy is a more serious possibility. For example, Papineau (2002) argues that there are a surprisingly wide range of different equally natural candidate properties for "consciousness" that fit with our usage equally well, and therefore, according to him we have reason to believe that "consciousness" (even once disambiguated as "phenomenal consciousness") is a radically, or even *defectively*, indeterminate term.

However, it could be that a view like Epistemicism is true (see e.g. Williamson (1994)), on which there is no semantic indeterminacy and use is always sufficient to fix a determinate referent for every term (even if we can't figure out what it is). So even if Deflationary Pluralism is true, "consciousness" could be a sharp term. Equivalently, a question that isn't substantive could still be one that has a determinate answer (here I am echoing Sider's (2011) discussion of this distinction). In general, the metaphysically interesting question is always whether a question is substantive, not whether it has a determinate answer – the indeterminacy question may at best combine the substantivity question with further questions about how words get their meaning, questions which aren't directly about the language-independent world.

---

[9] In most circumstances, if we have a case of vagueness that involves a sorites-like series of cases along which it is unclear where the extensional boundary of a predicate lies, we know how the application of the predicate depends on lower-level features (e.g. we know how baldness depends on hair distribution), and we can tell where the vague cases are : grasp of the predicate involves reacting in a "vagueness appropriate" way to the vague cases. Not so in the case of "conscious", if we believe in an inferential gap. We could have a creature in a state that is vague case of consciousness, and know all the facts about the case on which the application of "conscious" depends (e.g. all the physical facts), but have wide range of different opinions about whether the creature is conscious. So if "conscious" is vague, it is a very special example of vagueness.

We've now spent a considerable amount of time trying to clarify the Deflationary Pluralist view (although I still haven't said much about what "significance" is in its various guises: more on that below). I now move to a discussion of how the view might be motivated.

3 : Motivating DP : Reductionism and the Epistemology of Significance

The kind of argument I want to focus on involves carefully thinking about what the correct epistemology is for the facts about the significance of a property, including the naturalness of a property, or its normative significance. My view is that taking a reductionist view could have important implications for this issue, especially in the case of naturalness. Not only does reductionism mean that consciousness (or determinate forms of consciousness) lacks a certain kind of heavyweight naturalness by not being a fundamental property; it also may support a certain epistemology of naturalness (and perhaps of other kinds of significance too), an epistemology which may in turn support a DP view.

Recall that the reductionist holds that consciousness has a complex real definition in terms of more basic properties. This reductionist position has an obvious, but nonetheless seldom noted epistemological consequence. Even if consciousness has a complex physical nature, introspection doesn't reveal to us its complex nature (here I am assuming that views like analytical functionalism, on which we have a priori access to the real definition of consciousness, are false). As a result, introspection leaves it opaque to us what consciousness really is. On the reductive view, the nature of consciousness, far from being "completely revealed" from a first-person perspective (as "Cartesian" intuitions might suggest), is utterly hidden. To know what property consciousness really is requires explicitly articulating its complex structure, and this can only be done by using objective "third-personal" descriptive resources.

This has the following, somewhat startling consequence – a being that completely lacked consciousness could be in a better position than us to know what consciousness is. If they know which complex property our word "consciousness" refers to, and we don't, then that is exactly the situation that obtains.

For our purposes, the important ramification is that arguably this hypothetical being is in a better position to assess the various kinds of significance that consciousness may have than we are. After all, we don't know what consciousness is, and they do – doesn't that better equip them to assess its significance? Even if we think it is reasonable to take our beliefs about the significance of consciousness as 'innocent until proven guilty', surely once we discover its complex nature, it would be reasonable to hope that these beliefs would be 3rd personally *ratified*. Furthermore, arguably if these beliefs *fail* to be ratified from this perspective, then we ought to reconsider them. (I'll defend this idea in more detail in a moment.)

If we accept this "demand for 3rd personal ratification", what outcome can we expect the process of ratification to have, given what we already know? I claim that we

already know enough to be skeptical about whether our beliefs attaching heavyweight significance to consciousness are likely to be ratified from a 3rd personal perspective.

Imagine, for example, a super-intelligent alien scientist looking at the beliefs and practices of both human beings and the non-conscious aliens I mentioned earlier. Imagine furthermore that these observers lack consciousness, and they never reflect "from the inside" on their own psychological lives, so that their conception of mental states is purely theoretical. If these impartial zombie observers have perfect knowledge of both our psychological architecture and that of the unconscious aliens, are they likely to conclude that there is some special glow attaching to us humans, and not to the aliens? Similarly, if they describe the functioning of the human brain at many levels of abstraction, are they likely to find a particular level that is outstandingly significant, the level at which consciousness resides? I think this is prima facie implausible (I will try to spell out in more detail why this is in a moment).

It is tempting to object that although the various kinds of special significance that attach to consciousness aren't accessible to our impartial zombie observer, *we* are in a position to appreciate them, given that we *have* consciousness. Prima facie, this misses the point of the demand for third personal ratification. The point of the demand is that, given that consciousness is a complex physical property, the impartial observer is in a *better* epistemic position vis a vis its significance, even though the observer may lack consciousness, and we have it. One could challenge this claim however – can more be said to justify the link between reductionism and the demand for 3rd personal ratification?

A theorist who rejects the demand for 3rd personal ratification is favoring what I call a "top-down" epistemology, rather than a "bottom up" epistemology of significance (a very similar distinction is made by (Johnston (1997) in his discussion of the practical significance of personal identity with Derek Parfit). A top-down theorist takes their common-sense convictions about the significance of consciousness (convictions they may have as a result of reflecting on consciousness "from the inside") as default justified, and argues that whatever physical property consciousness is identical with must be significant *because* it is identical with consciousness. By contrast, the "bottom-up theorist" thinks that our beliefs about the significance of a property are in principle revisable given reflection on the real nature of the property: "Property P doesn't have the kind of significance we would normally attach to consciousness; consciousness is property P; therefore, consciousness isn't as significant as we thought it is."

Notice that if the bottom up perspective is right, it gives us a further insight into the sense in which certain questions about presence of consciousness in problem cases might be empty questions. If the 3rd personal perspective has this kind of epistemic privilege, then a being who knows all the physical facts about how, say, a lobster functions, is in an epistemically ideal state for knowing everything that matters about the lobster. In so far as this is consistent with them *not* being able to know

whether or not the lobster is conscious, failing to have this knowledge could not be failing to know anything significant.

Also notice that if this is right, then it could be harder than it looks to find a non-question begging starting point in debates about the viability of a reductionist view of consciousness. Using as a premise a claim about the significance of consciousness whose justification is based on imagining "from the inside" the difference between conscious and unconscious beings could be using a premise that might be called into question if we accepted reductivism and the allegedly ensuing demand for 3rd personal ratification.

Admittedly I can imagine scenarios which if actualized would superficially appear to support the top-down epistemology. For example, in his "Dark Materials" Trilogy Phillip Pullman imagines the existence of a golden dust that is in some sense the basis for consciousness [10] (Pullman is thinking of consciousness as something like self-consciousness – let's imagine that it is phenomenal consciousness instead). Our hero Lyra discovers that this dust is escaping from the world, and struggles to prevent its loss. Of course, what motivates her quest is the assumption that the dust is highly significant *because* it is the basis for consciousness. It would seem perverse for her to think : "hold on, if consciousness amounts to no more than the presence of some dust, why care about consciousness?"

The example is misleading however: the main problem is that it is hard to imagine the dust being anything more than the *causal* basis of consciousness, rather than literally constituting consciousness. Therefore, the story implicitly suggests that Property Dualism is true. If Property Dualism is true then consciousness is highly descriptively significant because it's a fundamental property, a fact that would be confirmed through application of a bottom-up epistemology. The dust would in turn be significant because of its direct causal-explanatory relationship with consciousness. The example is therefore a distraction in the context of a discussion where we are assuming reductionism.

Hawthorne (2007) objects to a bottom-up epistemology, on the grounds that it will inevitably deliver something like a Deflationary Pluralist view for many concepts we regard as significant. For example, one case he considers is the concept of *truth*. On a non-deflationary (in Field's (1994) sense) reductionist view of truth, it is a complex relation between sentences, thoughts and the like and worldly states of affairs. Presumably, there are small variations of its complex definition that correspond to nearby truth-like properties, such as "truth*". Hawthorne points out that we reasonably attach far more significance to truth over truth*, even though the justification for this is only accessible "from the inside" to those using the concept (e.g. I care whether it's true that my hair is on fire, not whether its true*, because truth* is consistent with falsity). He concludes that something like a top-down epistemology of significance must be correct for this concept, and other related

---

[10] Thanks to John Campbell for this example.

concepts like "belief" and "knowledge". One can imagine him extending the line to "consciousness" and other phenomenal concepts.

To fully address this worry would require looking in detail at each of these different concepts, to figure out how plausible a DP view would be in each case, which is beyond the scope of this paper. My general response is that it is hard to see how the bottom-up methodology could *fail* to be correct, given certain assumptions about the nature of significance. Specifically, if the significance of a property is a *function* of its complex real definition (or at least of that plus the facts about its context of instantiation that would be accessible to a suitably well informed zombie Martian observer, if significance is context-sensitive), it's hard to see how knowing the real definition wouldn't be an epistemically privileged position vis a vis significance (one alternative is that there could be an epistemic gap for significance without a metaphysical gap – more on that below). Furthermore, it is hard to see how the alleged "insight from the inside" into the significance of various properties is supposed to work; it is hard to see how to elaborate a "top-down" epistemology (this is a topic that I won't have any more to say about here).

This discussion suggests the following "master argument" for each form of deflationary pluralism:

(1) If reductive materialism is true, consciousness is a complex physical property.
(2) If it is a complex physical property, then it has a hidden nature that is not revealed to us from a first person perspective.
(3) The significance of a complex property is determined by its complex nature (perhaps in addition to the physical facts which provide its context of instantiation, if it is context-sensitive).
(4) Given (3), if reductive materialism is true, we are in a better position to assess the significance of consciousness from a 3rd person perspective.
(5) From a 3rd person perspective, each form of quasi-consciousness appears equally significant as consciousness, even though it is a different property from consciousness.
(6) Therefore (despite the conflict with 1st personal intuitions), if RM is true, we have reason to endorse deflationary pluralism.

Premise (1) is just an elucidation of how I'm understanding "reductive materialism", and Premise (2) ought to be accepted by anyone who rejects analytical materialism. I'll now spend some time defending (3), (4), and (5) focusing on a version of the argument where the relevant "significance" is the "naturalness" or "joint-carvingness" of consciousness. Is consciousness significantly more natural than other forms of quasi-consciousness?

<u>Defending Descriptive Deflationary Pluralism</u>

Some readers will be skeptical about whether there is any notion of objective naturalness that succeeds in distinguishing between properties (one can talk about

naturalness and fundamentality for other metaphysical categories such as facts and material objects, but properties will be our concern here). Following Goodman (1954), many people think that at best certain properties are particularly interesting to us, given our practical concerns, and the arbitrary ways in which we have chosen to formulate our theories. But as I understand it, this view is extremely radical – it says, in effect, that any arbitrary collection of actual and/or merely possible objects forms just as natural a grouping as any other, once we abstract away our human perspective. For example, there is no sense in which two electrons are objectively more similar than an electron and a banana. Thus understood, it is tantamount to the claim that the world is entirely homogenous and lacking in objective structure. This is a position that I find incredible. In this discussion I'll therefore follow Armstrong (1978) and Lewis (1984), and later authors in finding the claim that there exists some sort of objective distinction of naturalness between properties a reasonable theoretical starting point. The question then arises whether consciousness is distinctive in the naturalness that attaches to it (the reductive materialist denies that it is perfectly natural, but this leaves various options open).

Merely believing in some applicable notion of objective naturalness leaves a lot open. For example, is naturalness all-or-nothing, or is it best understood in graded or comparative terms (see Sider (2011) for a defense of the former view, and Schaffer (2009) for a defense of the graded view). There is much to be said about this issue – here I will limit myself to a few points that will be important in the present context. The core issue is whether the naturalness of a *high-level* property is a "further fact" about it, in the sense that it might not be accessible to our Martian observer, even given knowledge of the relevant physical facts, including knowledge of the real definitions of physical properties. If that were the case, consciousness could be much more natural than other physical properties with very similar real definitions (or distinguished by all-or-nothing naturalness), even though this fact is completely opaque to our Martian observer comparing us with an unconscious alien.

To be more precise, there are two ways in which there could be such an epistemic gap for naturalness. One is if the naturalness that belongs to high-level properties is *metaphysically* primitive, and therefore literally a further feature of the world, beyond what we are imagining is known to our Martian. Another view is that, although high-level naturalness is a derivative property, our concept of naturalness makes it opaque how the naturalness facts depend on the rest of the facts; just as in the case of consciousness itself, it makes sense to suppose that there might be an epistemic gap here without a corresponding metaphysical gap. So our Martian Observer, equipped with our concept of naturalness, wouldn't necessarily be able to know how natural consciousness is, despite knowing all the facts that are relevant to determining this.

At this point we need to ask : what are the different views one can have of the metaphysics of high-level naturalness – the kind of naturalness that belongs to consciousness on a reductive materialist view?

One view that should be noted for our purposes is a version of the all-or-nothing view that I'll call "High-level Nihilism". According to the High-level nihilist, there is an objective distinction between a base of fundamental physical properties and the rest, but there are no objective distinctions of naturalness between non-fundamental properties – there are no "real patterns" in the high-level world, consisting of the instantiation of combinations of relatively natural high-level properties. If consciousness is a complex high-level property, then this view obviously implies a highly deflationary take on its descriptive significance. Here I will assume that high-level nihilism is mistaken, and want to consider the implications of alternative views for our argument for DP.

There are two kinds of "all-or-nothing" views that allow for high-level naturalness. On a view like Armstrong's (1978), the "natural" properties are not just a base of fundamental properties that provide a determination base for everything else, but also include high-level properties like "tree" and "human", which are distinguished by being natural from properties like "grue". I regard this view as implausible, and also, assuming that its proponent would classify some phenomenal properties as primitively natural, not really consistent with reductive materialism as I'm understanding it. So here I'll just consider the other kind of "all-or-nothing" view, on which *only* a base of fundamental properties get to count as all-or-nothing natural (this is Lewis's (1984) view). On this view, high-level properties do get to count as more or less natural also, but only in a *derivative* sense (assuming that we are rejecting high-level nihilism). For example, on Lewis's version, the naturalness of a high-level property is its *length of definition* in terms of the perfectly natural properties. Both these views contrast with a *primitivist comparativist* view of high-level naturalness which takes a comparative notion of naturalness as primitive, perfectly natural properties simply being those at one end of a primitive scale (see Schaffer (2009) for a defense of this view).

Many find Lewis's derivativist view very implausible, on the grounds that many natural-seeming high-level properties will have very complex or even infinitary definitions, delivering the result that they are less natural than many bizarre seeming properties with more simple definitions. Hawthorne (2007), for example, sees the problems with Lewis's Derivativist view as a motivation for simply taking high-level naturalness as a free-floating primitive, not to be understood in other terms.

Despite this worry about Derivatist views, I'll now argue as follows:

(a) In so far as there is reason to take a notion of naturalness as primitive, we should understand it as an all-or-nothing notion that applies only to fundamental properties (and therefore not to consciousness, considered as a high-level property)
(b) High-level properties *can* be rated for degree of naturalness, but only in a derivative way (this is equivalent to premise (a) + the rejection of high-level nihilism).

(c) Derivate degree of naturalness is best understood as causal-explanatory significance.
(d) The causal explanatory significance of a high-level property can be determined from knowledge of the fundamental physical facts (or perhaps a more inclusive set of facts)– that is, it can be determined by our impartial zombie observer.
(e) The impartial zombie observer will not regard consciousness as more explanatorily significant than other consciousness-like properties.

Against the idea that high-level naturalness is a free-floating metaphysical primitive (which I take to be best represented by a primitivist comparativist view), there are a number of things to say.

First, some will find this view to be to be in too much tension with the spirit of the reductionist view we are taking for granted to really take seriously in this discussion.

Second, if high-level naturalness floats free of the definition of a property, then a serious epistemological problem emerges. If our hyper-informed Martian can't see the relative naturalness of properties, why think that we can? In the same vein, the primitivist view adds an interesting twist to the epistemological difficulties associated with the underdetermination of unobservable facts by observable facts. On a primitivist view, even once we fix the fundamental physical facts, there are still many epistemic possibilities for what the free-floating high-level patterns in the world are: we get a surprising kind of underdetermination in the micro to macro direction as well. This in turn adds a surprising extra dimension of uncertainty to inference *from* macro *to* micro.

Third, there is a worry about the relationship between high-level naturalness and causal explanation. We presumably want to hold onto the idea that the suitability of a property to enter into causal explanations is closely related to its naturalness. But high-level causal patterns are surely derivative from the distribution of fundamental facts; they aren't a primitive further feature of the world inscrutable from the base. It seems that the degree primitivist may have to either give up the naturalness / explanation connection, or adopt a kind of strongly anti-reductionist view of high-level causal patterns.

A fourth worry - if high-level naturalness is primitive, then arguably the facts about it would be completely *precise* (vagueness is not out there "in the world"); but it is hard to believe that there is a precise facts of the matter about the relative naturalness of say, being a labrador and being a rhododendron. Primitivism arguably introduces a kind of arbitrariness that is hard to accept. Better to think that high-level naturalness is a somewhat vague derivative notion, one that perhaps could be precisified in a number of equally good ways.

All this makes me think that although we don't know how to pull off the trick, to avoid high-level nihilism in a plausible way, a consciousness reductionist should

reject a Primitivist view and accept that high-level naturalness is derivative, even if not in the way that the simple Lewisian "naturalness as length of definition" view claims.

So, assuming that Lewis's length of definition view isn't much of an improvement on high-level nihlism, what alternative theory of derivative high-level naturalness could there be? As I just noted, our main grip on high-level naturalness seems to come from relating it to causal explanation. It is therefore plausible to think of it as just *defined* as causal explanatory significance.

Now admittedly, despite significant advances elucidating the content of causal concepts in terms of their relations to counterfactuals and probabilistic concepts, it is very unclear *how* to define high-level naturalness in causal-explanatory terms. This is especially clear if it is part of the task to explain how to derive the facts about high-level causal patterns from more fundamental facts. In fact, there are serious objections to the idea that this could in principle be done[11]. Nonetheless, since I'm strongly inclined to believe that high-level nihilism is false, and that this is the most plausible way to resist it, I remain optimistic.

Let's suppose then, that we have a grip on how there could be objective causal-explanatory patterns that are determined derivatively by the fundamental physical facts. According to premise (d), these facts, and therefore the relative causal explanatory significance of different high-level properties, are available to our Martian on the basis of knowledge of the fundamental physical facts. The plausible claim here is that causal-explanatory significance is *not subject to a relevant kind of epistemic gap*. It is plausible that if our Martian observer knows all the base level physical facts, they *are* in a position to compare the relative causal-explanatory significance of consciousness with other consciousness-like properties.

This may appear to be akin to the kind of strongly rationalist claim recently put forward by Chalmers and Jackson (2001), to the effect that all facts about the world can be derived a priori from knowledge of a base of fundamental facts (perhaps including the addition of "centering" information, to deal with knowledge of some quasi-indexical facts, such as facts about individuals or natural kinds). But for our purposes, it is far from clear that anything this strong is really required. The intuitive idea that needs to be captured is that the significance of consciousness is optimally accessed from a 3rd person perspective, where "significance" is now being

---

[11] A serious worry with equating high-level naturalness with causal-explanatory significance comes from the thought that the causal-explanatory significance of a property is always relative to an explanatory target, and that there is no objective sense in which one causal explanatory target is more objectively interesting or worth explaining than any other (notice that appealing to the target's own causal-explanatory significance here will apparently just start a regress). I won't respond to this worry here, merely noting that if it really has force, it threatens us with high-level nihilism again.

understood as causal-explanatory significance. This seems to be consistent with the base including not just information about the fundamental facts, or the relevant derivation not being strictly a priori.

The only question remaining then, is: would our Martian think that consciousness has outstanding causal-explanatory significance?[12] More specifically, I take that the question is : does consciousness play an outstandingly significant role in causally explaining our behavior, or any other important aspect of our lives, such as our formation of beliefs and desires?[13]. A truly satisfactory answer would involve looking at the complete story about what actually happens in people's brains when they have conscious experiences, a story of which we only have the sketchiest outline. It is therefore beyond the scope of this paper to address the question this way. I'll have to limit myself to what I take to be some key points of a more abstract kind, which I hope will persuade that the relevant deflationary pluralist view is at least *plausible*.

The background assumption I'll make here is the widely held view, elaborated by a number of authors including Block (1995), that explanations in empirical psychology are *functional analyses*. The brain can be functionally analyzed at many levels of detail (including at the neurophysiological level) and psychological similarity, in the relevant sense, consists in sharing a functional description. If we think of psychological properties or natural kinds as those properties in virtue of which systems are psychologically similar, then on this picture we enjoy very many psychological properties: a set of inter-defined functional properties will exist at each of the different levels of functional description that can be given of us.

This suggests a kind of Deflationary Pluralism for type (2) properties : the view that different levels of description of our cognitive architecture are not such that one level – the level at which one can specify necessary and sufficient conditions for consciousness – has special causal explanatory significance. The reason this has

---

[12] Admittedly, it is at best unclear whether our intuitions about the significance of consciousness are very well represented by the claim that it has a high degree of causal-explanatory significance. Speaking personally, I would say that this is not what I had in mind; even if consciousness was completely epiphenomenal, it might still have the kind of significance I was thinking about. So it's not clear that maintaining that (e) is true is even inconsistent with anything common-sense has to say about consciousness. Still, if the only viable defense against deflationary pluralist view of the naturalness of consciousness is in terms of its causal-explanatory significance, then denying (e) might be the only way to *indirectly* defend common sense.

[13] Of course, if these things are themselves cosmically insignificant, having this relation to them may not support the view that consciousness is objectively significant – see footnote 9.

prima facie plausibility is as follows. More abstract functional descriptions, and the corresponding functional properties, tend to have more explanatory generality, encompassing more instances, telling us the broad architectural details of the system. As we fill in more detail, we lose this generality, but we gain much insight from the increase in information about how the system functions. It seems doubtful that there's any fact of the matter about how to trade off these different explanatory advantages and disadvantages, nor that they will weigh up in such a way that one level of description will stand out. Without actually looking at the details of human psychology, this is the main reason to be suspicious in advance of the claim that consciousness is outstandingly explanatory.

For type (3) properties, the basic idea of the deflationist will be something like this : assuming that Superficialism fails, then an unconscious alien or robot could have states that share some of the causal-explanatory features of conscious states, without themselves being conscious states. It's implausible to think that the relevant internal states of the alien have much less explanatory significance wrt to *their* internal life and behavior than conscious states have for *our* behavior.

Of course, there is undoubtedly far more of interest to be said about the explanatory role of consciousness vs that of other similar properties: in many ways, these remarks are embarrassingly brief. At this point, I want to focus on an important objection to this line of thought that focuses on one particular way in which explanations involving consciousness might be special. This will involve connecting the causal-explanatory significance of consciousness with its epistemic significance, an important topic for further discussion.

It might be objected that if we are thinking of behavior as mere bodily movements, then it is pretty clear that there is no level of analysis of our cognitive architecture that is privileged for explaining these movements, for the reasons I just gave. But suppose we switch our attention from mere bodily movements to intentional actions, and the kind of explanation we are interested from ordinary mechanistic causal explanation, to intentional causal explanation: explanation in terms of content-bearing states like beliefs and intentions. Perhaps the fact that certain objects are intentional systems is in some way an objectively significant property of them, and the "level of description" which corresponds to picking out their intentional states can reasonably be thought to stand out as special. Moreover, suppose it's true that *consciousness has an ineliminable role to play in intentional explanation*, so that only conscious beings can correctly have their behavior explained in this way. That would plausibly make consciousness genuinely special.

Let me contrast two different ways of developing this picture, one of which I think is more promising than the other. On the pure functionalist picture, intentional states, including conscious states, are partly individuated by their functional roles, and intentional explanation is really just a species of purely causal explanation. The main explanatory role that distinguishing mental states by their contents is playing, is to give us information about their causal powers, powers which then come into play in explaining a subject's behavior or changes in their mental states. If this is all

there is to intentional explanation, then it's hard to see how there's anything particularly distinctive about the states that are involved in it (even if there is something distinctive about the way we pick them out in terms of contents), or, more importantly, why consciousness would be needed in order for states that can play this kind of causal role to exist. This latter point is illustrated by the fact that we already have a practice of assigning contents to the states of completely unconscious subsystems of the brain, a practice that can reasonably be interpreted as involving a purely functional kind of intentional explanation of this kind.

A contrasting picture is the epistemic picture, according to which intentional explanation is not purely causal, but also involves explaining why the relevant behavior or mental states are *reasonable* or *rational*: it has a normative component. This would make it a genuinely distinctive kind of explanation. Furthermore, if it were true that consciousness is required for this kind of rationalizing explanation to be applicable, then that might explain what is distinctively special about consciousness. There is even a prima facie case that consciousness is required for this kind of epistemic explanation: we do not think that it is appropriate to apply epistemic notions like reasonability or rationality to subpersonal states that are remote from consciousness, even if we attribute contents to them and invoke them in a form of intentional explanation. One way to justify withholding rational assessment from subpersonal states would be to appeal to a link between phenomenal consciousness and rationality (see Smithies (ms.) for a forceful defense of this position).

This response moves us away from the idea that consciousness is special because it has a high degree of explanatory significance in a purely causal sense, to a view on which it is special in an epistemic sense. Is there a sense in which conscious experience is outstandingly epistemically significant, or is a form of Deflationary Pluralism the correct view in this area as well? Unfortunately, I don't have space here to adequately address this question, and will have to limit myself to a few brief remarks about how I see the discussion continuing at this point (see Lee (manuscript) for a more detailed discussion of the epistemic significance of consciousness).

There are a number of different senses in which conscious experience might be thought to have special epistemic significance. Here I'll briefly discuss just one of them : the idea that there are certain types of beliefs – say, perceptual beliefs or introspective beliefs – that can only be justified if they are based on a conscious experience. For example, it might be held that there are objective epistemic norms governing the formation of perceptual beliefs, that imply that only perceptual beliefs formed be endorsing the contents of phenomenally conscious perceptual experiences can be justified. If all is dark within, you cannot really be justified in believing anything about your environment. Against this, one might claim that an unconscious alien with consciousness-like perceptual states could in principle have perceptual beliefs that were equally justified as ours – they could be "epistemically quasi-conscious".

How might one argue for the existence of epistemic quasi-consciousness? Clearly our unconscious alien could form perceptual beliefs that are just as reliable as ours. But we should note that we needn't lean on a crude form of reliablism about justification to argue that the alien has justification. There will be many internal structural similarities between us and the alien, and so many theories of justification with an internalist component (i.e. one that requires a justified belief to stand in certain internally specifiable relations to other internal states) may well be satisfied by the alien's beliefs also. It looks like rejecting the view that our alien is justified will require a brute appeal to consciousness as an essential epistemic ingredient, not merely a rejection of reliabilism about justification.

Another point is that an opponent of epistemic quasi-consciousness will have to make a case that they aren't mistaking local for global significance: it may be that *for us humans*, conscious experience is required for justification – for example, in close worlds where I don't consciously experience a blue cup in front of me, I'm not justified in believing such a cup is present. Could there be an illegitimate slide from a correct local claim of this kind, to an ill-founded global claim?

More generally, what exactly would justify believing that consciousness is essential for justification? Arguably, the opponent of epistemic quasi-consciousness will have to claim that it is simply *inconceivable* that an unconscious being could have justified perceptual or introspective beliefs. But there appears to be no incoherence or conceptual confusion in the idea of epistemic quasi-consciousness, or at least so I would argue. Perhaps more importantly (and as I mentioned earlier) one could argue that our ordinary view is that the epistemic significance of consciousness is grounded in its natural significance: i.e. conscious acquaintance with the external world and our own mental lives is a completely different kind of natural phenomenon from zombie acquaintance. If there is no deep natural difference here though, that puts pressure on these reasons for believing in an important epistemic difference.

To sum up : the naturalness of a high-level property is best understood not as a primitive further fact about it, but as a derivative fact, in particular it should be understood as its causal-explanatory significance. It seems plausible that our detached Martian would be able to compare the explanatory significance of consciousness with instances of pseudo-consciousness, and would, at least in some cases, regard them as equally significant – or so I argued. I considered the objection that there is a special kind of intentional explanation of our behavior that involves rationalizing it in terms of intentional states, and that only a conscious being could correctly have their behavior explained in this special way. This opens the door for further discussion of the role of conscious experience in epistemology; I briefly mentioned some reasons for doubting the claim that conscious experience is essential for certain beliefs such as perceptual beliefs to be justified.

This ends my positive case for a form of Deflationary Pluralism. I conclude by mentioning a few ways in which what I've said invites further discussion.

Connections and Questions for Further Discussion

What impact does Deflationary Pluralism have on other debates about consciousness, for example the debate about the alleged Hard Problem of consciousness, and the philosophical issues surrounding the empirical search for neural correlates of consciousness?

On the latter issue, DP may give us an extra reason to be skeptical about the resolvability of certain empirical questions about the physical basis of consciousness : it rules out a view on which one consciousness-like physical property is much more natural than others in a way that is in principle accessible to suitably well-placed scientific observer: there will be no ethereal glow attached to one property, telling us that it is the elusive "factor X" (I'm sure this is obvious to most people, but it is worth making explicit).

Related to this, if some form of DP is true, then certain questions about the physical basis of consciousness (or determinate forms of consciousness) could be less substantive than they appear. For example, suppose we have a group of consciousness-like physical or functional properties, and we are trying to decide which one *really is* consciousness. Perhaps this isn't a substantive question and should be ignored. Interestingly, conclusions like this might in turn might support a perspective on which the real problems understanding consciousness are more philosophical than empirical. For example, even if the question "which of these physical / functional properties is identical with consciousness?" is to some degree empty, there might remain a legitimate *philosophical* question – how could *any* physical or functional property - including these properties - be identical with consciousness?

Finally, let me mention a couple of ways in which DP impacts the debate about the hard problem of consciousness. First, some materialists who accept the existence of an inferential gap have tried to defend their view by arguing that they can account for our epistemic situation vis a vis consciousness in completely materialistically acceptable terms, without being able to bridge the gap. Chalmers (2006) argues against this approach as follows: (1) even if materialism is true, zombies are still conceivable, (2) zombies are not in the same kind of epistemic situation as us vis a vis their experiences. Conclusion : the materialist can only explain our epistemic situation by explaining why we aren't zombies, which would seem to require closing the (apparent) inferential gap. As Chalmers himself concedes, the argument assumes that when we conceive of a zombie, we conceive of a being that is in an epistemically impoverished state compared with us. But if epistemic DP is true, some zombies are epistemically quasi-conscious, so these two conceptions don't necessarily go together. Chalmers does try to finesse this problem, but in my view more discussion is needed here.

Second, I think there is a strand of the explanatory gap that can be deflated once we accept a deflationary pluralist view. One thing that might make a physical / functional property seem puzzlingly ill-suited to being the basis of consciousness is a kind of "specialness mismatch" intuition – any physical or functional property is likely to be obviously just one of a large family of properties that are similar in one respect or other, and it might seem completely arbitrary that it is this property rather than that one that is the elusive "factor X" that is necessary and sufficient for consciousness. The Deflationary Pluralist is in a position to treat this problem – they can explain away the apparent mismatch by downgrading the specialness of consciousness.

To be clear, I certainly don't think that this completely removes or deflates the explanatory gap. There will remain the fact that we can imagine factor X occurring without consciousness, however much knowledge we have of our physical constitution and relations to our environment. And it will still be true that there will be positive reasons for finding it unintelligible how conscious states could be just the same as any physical/functional states. Still, I believe that the interest of Deflationary Pluralism does extend to shedding light on the cluster of issues that arise from the idea that consciousness is subject to an explanatory gap.

References

Armstrong, D. M., 1978, *Universals and Scientific Realism*, vols. I and II, Cambridge: Cambridge University Press.

Block, N. 2002. 'The Harder Problem of Consciousness'. *The Journal of Philosophy*, Vol.XCIX, 8: 391-425.

Block, N. 1995. The Mind as the Software of the Brain. In An Invitation to Cognitive Science, edited by D. Osherson, L. Gleitman, S. Kosslyn, E. Smith, and S. Sternberg. Cambridge, MA: MIT Press.

Chalmers, D. (2011). Verbal Disputes. *Philosophical Review*, 120:4, 2011.

Chalmers and Jackson (2001). Conceptual Analysis and Reductive Explanation. *Philosophical Review*, 110:315-61.

Chalmers, D. (2006) – Phenomenal Concepts and the Explanatory Gap. In (T. Alter & S. Walter, eds) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford University Press, 2006.

Dretske, F. (2000) – What Good is Consciousness? In Perception, Knowledge and Belief : Selected Essays. Cambridge University Press.

Field, H.(1986). 'The Deflationary Conception of Truth', in G. MacDonald and C. Wright (eds.), *Fact, Science and Morality*, Oxford: Blackwell.

Goodman, N. (1954) – Fact, Fiction, Forecast. Harvard University Press.

Hawthorne (2007) – Craziness and Metasemantics. *Philosophical Review* 116 (3):427-440.

Johnston M. (1997) – Human Concerns without Superlative Selves. In Dancy (ed.) Reading Parfit. Oxford: Blackwell.

Lee (manuscript) – The Epistemic Significance of Consciousness.

Lewis, D. 1972. 'Psychophysical and Theoretical Identifications', *Australasian Journal of Philosophy*, 50, 249-58.

Lewis, D., 1983, "New Work for a Theory of Universals", *Australasian Journal of Philosophy*, 61: 343–77

McGinn, C. 1989. 'Can We Solve the Mind-Body Problem?', *Mind* 98:349-66.

McLaughlin, B. P. (2003), A Naturalist–Phenomenal Realist Response to Block's Harder Problem. Philosophical Issues, 13: 163–204.

Papineau, D. (2002) Thinking about Consciousness. OUP.

Parfit, D. (1984) – Reasons and Persons. OUP.

Parfit, D. (1995) – the Unimportance of Identity. In H. Harris (ed.), *Identity*. Oxford University Press.

Pullman (2003) – His Dark Materials (Trilogy of Novels). Laural Leaf.

Schaffer, J. (2009) – On what Grounds what. In *Metametaphysics*, eds. Chalmers, Manley, and Wasserman (2009), 347-83: Oxford.

Shoemaker, S. 1975. Functionalism and qualia. *Philosophical Studies* 27:291-315.

Sider, T. (2011) – Writing the Book of the World. OUP.

Smithies, D. (ms.) – The Mental Lives of Zombies.

Strawson (1994) – Mental Reality.

Williamson, T. (1994). Vagueness. Routledge.